TRAINING-PATH EQUIVALENCE: VALIDATING IMPLEMENTATION SUBSTITUTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Efficient neural kernels are typically validated through output and gradient agreement at fixed weights. Training introduces a different requirement: a replacement must preserve the reference behavior after repeated parameter and optimizer updates. We introduce Training-Path Equivalence (TPE), a framework that turns this requirement into an explicit substitution contract and uses short paired training paths to screen candidates. In a fresh evaluation of two previously frozen policies on 35 condition–seed units, the step-50 paired screen identifies all 20 contract rejections and all 15 passes; local checks at fixed numerical tolerances through step 200 miss three rejections and reject five passing units. A separately timed Math-to-Flash attention substitution, accepted in five fresh seeds under the lower-learning-rate 5,000-step contract, retains a $1.357\times$ full GPU training-step speedup. Matched reference-seed calibration across nine conditions independently reproduces rejection and acceptance outcomes and quantifies the accepted pair's small functional gaps. Complementary long-horizon studies, replications up to 1.003B parameters, and optimizer-state interventions characterize how substitution decisions depend on training requirements. Together, these results establish paired-path validation as a practical complement to local numerical checks for selecting efficient training implementations.

1 INTRODUCTION

Fused operators, efficient attention kernels, and low-precision implementations are central to modern neural network training (Vaswani et al., 2017; Ba et al., 2016; Chen et al., 2018; Paszke et al., 2019; Dao et al., 2022). Before adopting a replacement, developers compare outputs and gradients at a fixed parameter state. These checks provide a useful numerical qualification gate (Goldberg, 1991; Gupta et al., 2015; Micikevicius et al., 2018). The operational decision, however, concerns an entire training procedure: will the faster implementation preserve the behavior that matters under the intended optimizer, precision, and training horizon?

The target here is reference-preserving substitution: replacing an implementation while retaining specified behavior of an established training procedure. This matters when a replacement is expected to reproduce a reference checkpoint or preserve intermediate behavior used for debugging and validation. The contract makes these requirements explicit, so the evaluation can be tied to the behavior the application actually needs.

A fixed-state comparison captures only one part of this decision. During training, numerical differences enter parameter updates and optimizer states, changing the states on which subsequent computations act. Conversely, a detectable local discrepancy may remain compatible with the required training behavior. A useful validator must therefore distinguish both outcomes: substitutions that violate the training requirements and substitutions that preserve them while delivering an efficiency benefit.

We introduce *Training-Path Equivalence* (TPE), a reference-conditioned framework for making this decision explicit. A substitution contract specifies the implementation pair, training protocol, observables, tolerances, checkpoint aggregation, and horizon. Matched reference and candidate branches then provide the evidence for acceptance. Short paired prefixes support early screening, while reference-seed calibration contextualizes the magnitude of the observed gaps. Figure 1 illus-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

trates why the declared aggregation matters: preserving the final state and preserving every monitored checkpoint are different requirements.

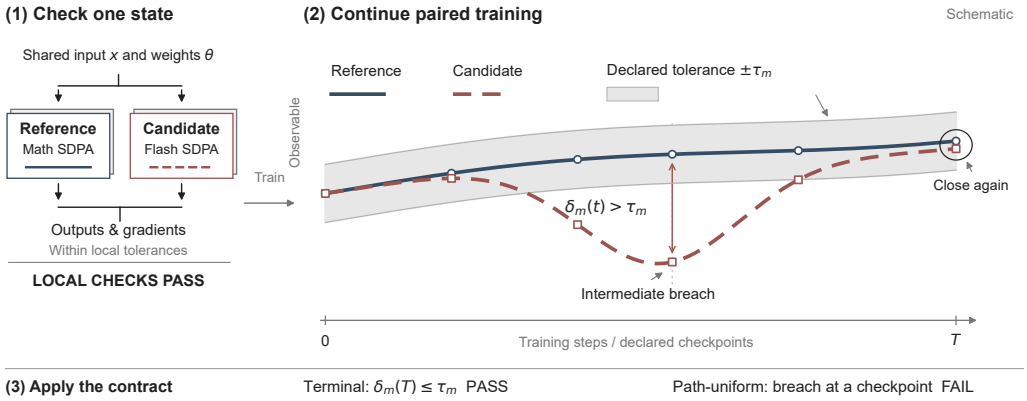

Figure 1: **Local checks do not certify the training path.** **(1)** The reference and candidate pass output and gradient checks at fixed weights. **(2)** During training, the candidate (dashed red) leaves the tolerance band around the reference (solid navy), then returns by T. **(3)** The terminal contract passes, but an intermediate declared checkpoint violates the path-uniform contract. The schematic band denotes an absolute-gap tolerance for one scalar observable, assuming all others pass; it is neither measured data nor a confidence interval.

Our central evaluation asks how two frozen validation policies agree with a declared training contract. On fresh confirmation seeds, the step-50 paired-path screen matches the final 5,000-step outcome on all 35 condition–seed units: 20 rejections and 15 passes. Reapplying the fixed numerical gate through step 200 misses three rejections and incorrectly rejects five passing units. In this cohort, the paired screen therefore resolves both types of error made by the specified local policy, supporting its use alongside numerical qualification.

The acceptance side has a concrete systems benefit. The public Math-SDPA/Flash-SDPA pair passes a lower-learning-rate 5,000-step terminal contract in five fresh seeds, and an independent timing study measures a $1.357\times$ speedup for a full GPU training step, including forward, backward, and optimizer update. A nine-condition study with matched reference calibration reproduces the acceptance and rejection outcomes. Broader experiments characterize the decision across horizons, model sizes up to 1.003B, and optimizer-state interventions.

Contributions. Our contributions connect a validation target, a decision procedure, and a measured systems benefit:

1. **A reference-conditioned training contract.** We specify what a replacement must preserve through the protocol, observables, tolerances, aggregation, and horizon. This formulation distinguishes terminal preservation from path-uniform preservation and makes an acceptance decision interpretable when training requirements change.

2. **Prospective validation of a paired-path screen.** A previously frozen step-50 rule correctly classifies all 20 rejections and 15 passes in a fresh cohort. It resolves cases missed or incorrectly rejected by fixed-tolerance local checks through step 200, demonstrating practical value relative to this numerical qualification policy.

3. **Validated efficiency with complementary diagnostics.** Independent confirmation accepts a non-exact Math-to-Flash substitution that retains a $1.357\times$ full GPU step speedup. Matched seed calibration contextualizes its functional gaps, while long-horizon studies, replications up to 1.003B parameters, and native optimizer-state interventions characterize the conditions governing substitution.

2 TRAINING-PATH EQUIVALENCE

2.1 STATE EVOLUTION AND REFERENCE-CONDITIONED CONTRACTS

Let $i \in \{r, c\}$ index a reference and a candidate intended to compute the same function. At a shared parameter state, write the recorded local statistics on audit batch B_j as $a_j = (e_j^{\mathrm{out}}, e_j^{\mathrm{grad}}, c_j^{\mathrm{grad}})$: output relative maximum error, gradient relative ℓ_2 error, and gradient cosine. The frozen four-batch qualification gate is

$$Q_{\mathrm{loc}} = \prod_{j=1}^{4} \mathbf{1}\Big[e_j^{\mathrm{out}} \le 10^{-2},\ e_j^{\mathrm{grad}} \le 2\times 10^{-3},\ c_j^{\mathrm{grad}} \ge 0.999\Big]. \tag{1}$$

Exact controls and non-exact candidates use the same gate. Training additionally evolves optimizer memory. With augmented state $z_t^{(i)} = (\theta_t^{(i)}, s_t^{(i)})$, shared initialization z_0, and a matched batch sequence B_t, define

$$z_{t+1}^{(i)} = F_{i,t}(z_t^{(i)}) = U_i\Big(\theta_t^{(i)}, s_t^{(i)}, G_i(\theta_t^{(i)}; B_t)\Big), \qquad G_i(\theta; B) = \nabla_\theta \ell(f_\theta^{(i)}; B). \tag{2}$$

The protocol fixes the runtime and evaluation; each branch retains its own state. U_i can itself be the substitution, as in fused versus unfused AdamW. The maps are conditioned on the realized execution; this notation does not assert runtime determinism.

Contract functional. A contract $\mathcal{C} = (r, c, \mathcal{P}, \mathcal{M}, \mathcal{A}, \boldsymbol{\tau}, T)$ fixes the protocol, observables, checkpoint aggregations, positive tolerances, and horizon. For an observable $O_{m,t}$ with declared nonnegative discrepancy d_m, let

$$\delta_m(t) = d_m\Big(O_{m,t}(z_t^{(r)}), O_{m,t}(z_t^{(c)})\Big), \quad g_m = A_m(\{\delta_m(t) : t \in \mathcal{S}_T\}), \quad R_{\mathcal{C}} = \max_{m\in\mathcal{M}} \frac{g_m}{\tau_m}. \tag{3}$$

Here $\mathcal{P}$ includes model, runtime, precision, optimizer, initialization, data order, and evaluation; $\mathcal{S}_T$ is the declared checkpoint grid containing T. Reference-normalized discrepancies need not be symmetric. The acceptance functional is

$$\mathrm{Accept}(r \to c \mid \mathcal{C}) = Q_{\mathrm{loc}}\, \mathbf{1}[R_{\mathcal{C}} \le 1], \qquad g_m = \begin{cases} \delta_m(T), & \text{terminal,} \\ \max_{t\in\mathcal{S}_T} \delta_m(t), & \text{path-uniform.} \end{cases} \tag{4}$$

The local gate applies at the states declared by the contract. Path-uniform acceptance implies terminal acceptance on the same grid; the converse need not hold. A contract profile evaluates this functional as tolerances, aggregation, or horizon vary (Appendix K.1). All confirmatory choices are frozen before evaluation.

The raw observables are parameter relative distance, $\max\{0, 1 - \mathrm{hidden_CKA}\}$, and absolute validation-loss gap, with tolerances $(0.05, 0.05, 0.001)$. Their checkpoint rule is

$$\mathrm{reject} = [\mathrm{param_rel} > 0.05] \vee [\mathrm{hidden_CKA} < 0.95] \vee [|\mathrm{val_loss_gap}| > 0.001]. \tag{5}$$

These tolerances were fixed through pilot-scale exact and SGD controls. They encode reference preservation: finite, learning runs may be rejected without worse predictive quality. Applications concerned only with predictive utility can declare functional observables. Logit and prediction disagreement supply additional diagnostics; the original stress separation persists at $0.5\times$ and $2\times$ tolerances (Appendix Table 32).

2.2 FROM LOCAL DEFECTS TO ACCUMULATED PATH DISCREPANCY

Define the state separation $e_t = z_t^{(c)} - z_t^{(r)}$ and the same-state update defect $\eta_t = F_{c,t}(z_t^{(r)}) - F_{r,t}(z_t^{(r)})$. Adding and subtracting the candidate update at the reference state yields the exact decomposition

$$e_{t+1} = \underbrace{F_{c,t}(z_t^{(c)}) - F_{c,t}(z_t^{(r)})}_{\text{propagation of the existing state separation}} + \eta_t. \tag{6}$$

If, along the compared states, the first term has norm at most $L_t\|e_t\|$ and $\|\eta_t\| \le \varepsilon_t$, then shared initialization gives

$$\|e_T\| \le \sum_{k=0}^{T-1} \varepsilon_k \prod_{j=k+1}^{T-1} L_j. \tag{7}$$

Thus even a bound on the complete update defect requires propagation control to certify a path. This conditional inequality motivates the measurements; the experiments do not estimate L_t or use it as an acceptance certificate. Appendix K.2 proves the bound and gives a scalar construction in which uniformly small local defects accumulate beyond a terminal tolerance.

For the actual gradient audits, the corresponding identity on a shared batch is

$$G_c(\theta_t^{(c)}; B) - G_r(\theta_t^{(r)}; B) = \underbrace{G_c(\theta_t^{(r)}; B) - G_r(\theta_t^{(r)}; B)}_{\text{same-state implementation difference}} + \underbrace{G_c(\theta_t^{(c)}; B) - G_c(\theta_t^{(r)}; B)}_{\text{difference induced by the evolved states}}. \tag{8}$$

Re-anchored local checks observe the first term; paired paths observe their sum after separate parameter and optimizer-state updates. The terms may reinforce or cancel. Repeating a same-state check therefore need not reproduce a decision about accumulated behavior. E2 tests this distinction with frozen gates and a common final contract.

2.3 Frozen Screening and Matched Calibration

Let $\rho_t = 1 - \mathrm{grad_cosine}_{\mathrm{pair}}(t)$ be the recorded gradient disagreement between branches at their own states. The prospective screen predicts raw-contract rejection through

$$\widehat{y}_{50} = \mathbf{1}[\rho_{50} > \gamma], \qquad \gamma = 1.8477439880371094 \times 10^{-5}. \tag{9}$$

The threshold was selected on earlier TinyStories training rows and is unchanged in E1–E2. The screen prioritizes evaluation; full paired training determines the outcome $y_{\mathcal{C}} = \mathbf{1}[R_{\mathcal{C}} > 1]$ in the locally qualified cohort.

For each fresh condition, five independent-initialization reference runs supply ten pairwise gaps. Write $g_m^{r_a,r_b;\mathrm{probe}}$ for their aggregated discrepancy and $g_m^{r,c;\mathrm{probe}}$ for the confirmation numerator, with protocol, horizon, aggregation, and common probe matched. Then

$$b_m = \binom{K}{2}^{-1} \sum_{1 \le a < b \le K} g_m^{r_a,r_b;\mathrm{probe}}, \quad K = 5, \qquad D_{\mathrm{TPE}}^m = \frac{g_m^{r,c;\mathrm{probe}}}{\max(b_m, 10^{-12})}. \tag{10}$$

Each denominator averages ten dependent comparisons from five runs; it measures initialization variation, not a coverage bound or same-seed repeatability. Common-probe ratios do not relabel the legacy-evaluation raw outcomes. Appendix A details matching; Figure 3 gives the workflow. Screening evaluation retains every final cohort outcome to count both error types; independent timing measures benefit under the confirmed workload.

3 Experimental Design

We study Transformer language models on TinyStories and WikiText2 (Vaswani et al., 2017; Eldan & Li, 2023; Merity et al., 2017). The base model has 477,696 parameters, width 128, two layers, four heads, byte vocabulary 256, sequence length 128, and batch size 16. Five-seed confirmations run for 5,000 steps with BF16 autocast, FP32 parameters and native optimizer state, matched initialization and data order, and AdamW with betas $(0.9, 0.999)$, epsilon 10^{-8}, and weight decay 0.01. Learning rates are constant: 3×10^{-3} by default and 10^{-4} for the acceptance protocol selected in an earlier tuning-only study.

Implementation pairs. **A0** compares public PyTorch Math and Flash SDPA; **A1**, explicit attention and Flash SDPA; **N0**, native and composed LayerNorm with matched epsilon; **O0**, native unfused and fused AdamW; and **T0**, matmul/einsum attention that is exact on the evaluated runtime. L8/L4 vary LayerNorm affine-gradient reduction order. Appendix J.1 specifies the calls and additional candidates.

Table 1: **Two frozen validation policies (E2).** Positive denotes rejection by the 5,000-step contract. All rows use the same 35 condition–seed units: 20 rejections and 15 passes. Local checks retain fixed numerical tolerances; the paired screen retains its earlier calibrated threshold.

Method / observation	Correct reject	Correct accept	False reject	Missed reject
Initialization local / 0	0	15	0	20
Repeated local / 50	12	15	0	8
Repeated local / 100	16	10	5	4
Repeated local / 200	17	10	5	3
Paired path / 50	**20**	**15**	**0**	**0**

Fresh confirmation and timing. The added study freezes nine conditions on an RTX 4090 with PyTorch 2.6.0+cu124: default A0/N0/O0/T0 on both datasets, plus WikiText2 A0 at the lower learning rate. Each condition uses five reference calibration seeds and five disjoint confirmation seeds, totaling 45 reference trajectories and 45 pairs (135 single-branch trajectories). E1 provides matched calibration and replication; E2 records local audits within the same confirmation trajectories; E3 separately times full GPU training steps using five ABBA groups per learning rate. E2 contains 35 condition–seed units from A0/N0/T0, with five seed IDs reused across conditions. O0 is excluded because an identical forward/backward computation is not an optimizer-local comparator.

Complementary studies. The earlier matrix includes exact and SGD controls, a 50,000-step base-model extension, hardware-stratified 355M and 1.003B replications, and native optimizer-state interventions. Their original protocols and cohorts remain separate from E1–E3. Appendix Table 34 and Appendix A provide configurations, seed roles, checkpoint schedules, and cost accounting.

4 Paired-Path Screening Beyond Fixed-Tolerance Checks

4.1 Comparison with a Fixed-Tolerance Local Policy

E2 evaluates repeated numerical qualification: reapplying the local gate at successive training states. Its tolerances remain fixed to test this policy's agreement with the final contract. At steps 0, 50, 100, and 200, both implementations are re-anchored to the *same reference parameter state* and audited on four fixed batches. The local policy rejects after any failure; its step-50, step-100, and step-200 decisions use 8, 12, and 16 checks. State and RNG audits verify that these checks preserve the live trajectories.

Table 1 compares both methods against the frozen 5,000-step raw outcome. Initialization checks accept every unit. Repeating local checks recovers some rejections, but at step 200 it still misses three and incorrectly rejects all five lower-learning-rate A0 units. The frozen step-50 paired-path rule identifies all 20 rejections and preserves all 15 passes. The five seed IDs are shared across conditions, so the 35 units describe this evaluated cohort rather than 35 independent seeds.

Why the decisions differ. By step 50, local checks identify all ten default A0 rejections but only two of ten N0 rejections. For WikiText2 N0 seed 34101, all 16 local batch checks through step 200 pass, whereas the step-50 paired gradient disagreement is 2.9800×10^{-5}, above the frozen threshold. Its final parameter distance is 0.6230 and hidden CKA is 0.8347, confirming rejection. Conversely, local failures in the lower-learning-rate A0 condition do not imply failure of its terminal contract. These cases show why both accumulated path differences and the required final behavior matter. Observation step indexes information availability; the methods have different computational costs (Appendix A.4).

Policy interpretation. The 15 passes comprise ten exact T0 controls and five non-exact lower-learning-rate A0 units. Excluding T0, the same screen correctly classifies all 20 rejections and five passes within the existing cohort. The local policy preserves a numerical qualification criterion, while the paired rule uses a threshold calibrated on earlier training rows to predict contract outcomes; both are frozen before confirmation. The policies retain different calibration objectives and

Table 2: Transfer of the pairwise screen under fixed AdamW. Features are selected on training rows; intervals are class-stratified bootstrap 95% CIs and AP is average precision.

Held-out target	Selected step-50 feature	AUC (95% CI)	AP	n_+/n_-	Local AUC	Metadata AUC
TinyStories	Hidden disagreement	0.841 [0.603, 1.000]	0.893	$9/7$	0.691	0.691
WikiText2	Gradient disagreement	0.900 [0.700, 1.000]	0.735	$6/10$	0.633	0.700
Dimension-512, six-layer	Logit symmetric KL	0.930 [0.804, 1.000]	0.937	$11/13$	0.696	0.710

Table 3: **Full GPU training-step timing (E3).** WikiText2 A0 on RTX 4090; five ABBA groups per learning rate. Times and within-group speedup ratios are summarized separately by their medians. Peak memory is CUDA allocated memory.

Learning rate	Math ms	Flash ms	Speedup	Math MiB	Flash MiB
3×10^{-3}	5.555	3.949	$1.348\times$	65.47	55.47
10^{-4} (accepted)	5.571	4.034	$\mathbf{1.357}\times$	65.47	55.47

computational costs. E2 therefore characterizes these predeclared workflows, with its comparative conclusion restricted to the fixed-tolerance baseline.

4.2 Transfer Across Earlier Held-Out Studies

Earlier fixed-AdamW tests evaluate feature selection on training rows and prediction on held-out pair–seed rows. Early-path AUC point estimates exceed local-only and metadata controls in all three tests (Table 2). Paired confidence intervals for the early-minus-local difference exclude zero on WikiText2 and dimension-512 transfer, while the TinyStories interval includes zero. The fresh repeated-local comparison above evaluates a single previously frozen rule rather than selecting a new feature on its test cohort.

Prospective tests also apply the frozen rule to locally qualified A0/A2/A3 candidates with 1.43–$1.53\times$ kernel speedups. A0 and A2 subsequently fail in all five long runs. Tuning-only learning-rate selection then yields an A0 acceptance without refitting the screen (Appendix Tables 14 and 15). This sequence makes the screen useful for a declared training configuration and motivates evaluating the efficiency of the accepted substitution.

5 Accepted Substitutions Retain Training Efficiency

5.1 Independent Acceptance and Full-Step Speedup

The same Math/Flash pair supports both rejection and acceptance under different training requirements. In the original lower-learning-rate confirmation, all five fresh seeds pass the 5,000-step terminal contract despite non-exact local differences: median parameter distance is 1.74×10^{-4}, hidden CKA is 0.999993, and validation loss decreases by 3.513. The fresh E1 cohort independently reproduces all five passes. On its matched common probe, median normalized gaps are 0.004439 for loss, 0.0002135 for hidden representations, 0.001251 for logits, and 0.02988 for predictions (Appendix Table 6).

E3 measures whether this accepted substitution retains a benefit at the training-step level. Five independent ABBA groups per learning rate execute Math, Flash, Flash, Math, resetting model and optimizer state before each segment. After 50 warmup steps, each segment measures 200 synchronized GPU steps, including forward, backward, and optimizer update. Data reside on the GPU; loading, host-to-device transfer, initialization, and warmup are outside the measured interval.

The accepted protocol retains a $1.357\times$ full-step speedup and reduces peak allocated memory by 10 MiB (Table 3). This result measures the benefit of selecting the existing Flash backend under a confirmed training contract. The original $1.618\times$ figure measures a separate attention-only forward/backward probe (Appendix Table 17). Each E3 speedup is the median of five within-group ratios, so it need not equal the quotient of the two displayed time medians. The accepted-row comparison holds learning rate fixed between Math and Flash. It measures implementation speed within

that protocol; it does not compare the time needed for different learning-rate protocols to reach the same validation quality.

Validation cost. The median step-50 paired prefix plus initialization audit takes 0.829 seconds under the acceptance protocol. The full five-reference calibration and five-pair confirmation cost approximately 392 seconds in summed recorded wall time, including the screen and audit. Using the short-segment timing to extrapolate savings over 5,000 steps gives approximately 51 reuses of this same workload to amortize that full validation cost. Appendix A.4 provides the default-rate counterpart and the accounting boundaries.

5.2 Matched Calibration Reproduces Both Decision Outcomes

Across the nine E1 conditions, default A0/N0/O0 reproduce 30/30 rejections, T0 reproduces 10/10 passes, and lower-learning-rate A0 reproduces 5/5 passes. The matched calibration strengthens the accepted case by placing its functional gaps far below the corresponding mean reference-seed gaps. It also clarifies the interpretation of rejection: default O0 is rejected by the raw contract even though its new loss and prediction ratios are below one on both datasets. These outcomes distinguish preserving a matched reference from remaining close relative to average inter-seed variation. The declared raw requirements determine substitution decisions, with fresh ratios providing a complementary diagnostic scale (Appendix A.2).

From validation to implementation choice. E1–E3 address complementary questions about the same substitution. E1 establishes which frozen conditions pass and measures their reference-relative gaps; E2 evaluates early screening against repeated checks at fixed numerical tolerances; E3 measures the training-step benefit of the accepted condition. Taken together, they support choosing Flash under the confirmed lower-learning-rate contract, with a recorded horizon, tolerance rule, and timing scope. The default-rate condition also benefits from Flash in the timing probe but is rejected by the preservation contract. Thus the framework supplies a concrete criterion for relating numerical behavior to an efficiency-driven implementation choice.

6 Characterizing Training-Contract Boundaries

6.1 Local Qualification and Long-Horizon Behavior

The earlier five-seed, 5,000-step matrix establishes the broader phenomenon across implementations. A0/A1/N0/O0 violate the default-AdamW contract on both datasets, while all 50 matched SGD runs and the exact T0 controls remain stable (Appendix Table 10). Independent L8/L4 experiments provide a second non-exact acceptance boundary: L8 passes in 5/5 WikiText2 runs at learning rate 10^{-4}, compared with 0/5 under the stress protocol (Appendix Table 19).

Figure 2 examines the A0 contract along additional coordinates. In the preregistered 50,000-step extension of the original accepted cohort, all five seeds pass at the final checkpoint; median parameter distance is 0.00379, hidden CKA is 0.99945, and loss gap is 7.10×10^{-4}. Each seed crosses the original loss-gap threshold at an intermediate checkpoint, however, so all five fail the path-uniform requirement. Stable counts at 5k/10k/20k/30k/40k/50k are 5/5, 2/5, 4/5, 4/5, 5/5, and 5/5. This transient crossing is precisely the distinction represented by the contract's aggregation.

The boundary is tolerance-sensitive: report-grid path-uniform passes are 0/5, 0/5, and 4/5 at $0.5\times$, $1\times$, and $2\times$ tolerances. At $2\times$, the denser supplementary grid yields 2/5 passes. Reporting both aggregation and checkpoint grid therefore gives a more informative account than a single unconditional stability label (Appendix Tables 26 and 27).

6.2 Hardware-Stratified Large-Model Replications

Real-kernel rejection persists in controlled larger-model studies (Table 4). All five 355M A0 runs on RTX 4090 and all five 1.003B A0 runs on RTX 6000D pass the local gate and violate the 5,000-step contract; exact T0 controls remain aligned over their declared horizons. The runs remain finite, showing reference-relative separation during training.

(a) Same pair, different contracts

Math-SDPA → Flash-SDPA

LR	Steps	Aggregation	Pass	Decision
0.478M / RTX 4090				
3×10^{-3}	5k	Terminal	**0/5**	**Reject**
10^{-4}	5k	Terminal	**5/5**	**Pass**
10^{-4}	**50k**	**Terminal**	**5/5**	**Pass**
10^{-4}	**50k**	**Path-uniform**	**0/5**	**Reject**
354.9M / RTX 6000D				
10^{-4}	20k	Terminal	**0/3**	**Reject**
1.003B / RTX 6000D				
3×10^{-4}	5k	Terminal	**0/5**	**Reject**

(b) Tolerance sensitivity

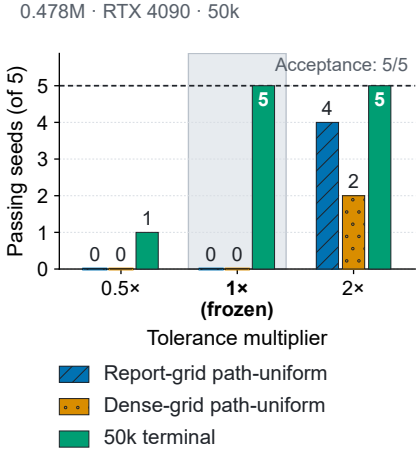

Figure 2: Contract profile for the public A0 Math-SDPA/Flash-SDPA pair. **(a)** Pass/reject decisions change with the declared learning rate, horizon, aggregation, model, and hardware/runtime stack; strata are not pooled. The highlighted 50k rows differ only in aggregation. **(b)** The 50,000-step path-uniform decision is tolerance-sensitive. At $2\times$ tolerance, 4/5 seeds pass the preregistered report grid and 2/5 pass the dense supplementary grid, while all five pass at the terminal checkpoint. All observed crossings are due solely to validation-loss gap.

Table 4: Real-kernel scale replications on WikiText2, reported separately by hardware. Rows show medians; seeds are never pooled across GPUs.

Pair	Steps	n	Div.	Param rel	Hidden CKA	Pred. agree
4090 355M A0	5,000	5	5/5	0.5727	0.7258	0.6406
4090 355M T0	500	3	0/3	0.0000	1.0000	1.0000
6000D 355M A0, LR 10^{-4}	20,000	3	3/3	0.4074	0.9367	0.8516
6000D 1B A0	5,000	5	5/5	0.6906	0.4595	0.7188
6000D 1B T0	1,000	5	0/5	0.0000	1.0000	1.0000

A separate 355M lower-learning-rate study evaluates three fresh seeds for 20,000 steps on RTX 6000D. All pass the local gate, first cross the unchanged raw rule at step 500, and fail the terminal and path-uniform contracts at step 20,000; median loss gap is 0.0260. This evaluates transfer to a combined model, runtime, and horizon setting and motivates revalidation when those coordinates change. It does not isolate the effect of model size. Independent V100 experiments further reproduce 355M O0/L8 rejection in 3/3 seeds per dataset; at 1.003B, O0 crosses the rule in 1/3 seeds by step 1,000 (Appendix Table 28).

6.3 Optimizer-State Interventions Explain O0 Amplification

O0 provides a controlled probe of how small implementation differences accumulate. The native fused and unfused AdamW branches have identical local forward/backward computations and a matched single-step relative update difference of only 1.74×10^{-5}–1.76×10^{-5}. We intervene after each native update by replacing both branches' first- or second-moment tensors with their elementwise mean while leaving parameters and gradients unsynchronized.

First-moment synchronization reduces median parameter distance by roughly an order of magnitude, raises hidden CKA to 0.9992, and reduces prediction disagreement from 0.17–0.20 to 0.013–0.017. Second-moment synchronization leaves substantial separation. For idealized AdamW, a first-order

expansion exposes gradient feedback through the first moment:

$$\delta m_t = \beta_1 \delta m_{t-1} + (1-\beta_1)\delta g_t, \tag{11}$$

$$\delta\theta_{t+1} \approx (1-\alpha\lambda)\delta\theta_t - \alpha\left(\frac{\delta\hat{m}_t}{d_t} - \frac{\hat{m}_t \delta d_t}{d_t^2}\right), \quad d_t = \sqrt{\hat{v}_t} + \epsilon. \tag{12}$$

Here λ is weight decay; Appendix K.3 gives the exact ratio identity and native rounding residuals. These interventions support a first-moment-dominant mediator within O0 (Appendix Figure 4 and Table 13). Increasing epsilon without matching update size also contracts paired gaps, but 18 independently calibrated update-norm-matched evaluations do not reproduce this contraction. The unmatched epsilon result is therefore scale-sensitive, while the native-state intervention provides the component-level evidence (Appendix Table 24).

7 Related Work

Numerical validation and reproducibility. Floating-point, seed, and tooling studies characterize reproducibility (Goldberg, 1991; Henderson et al., 2018; Summers & Dinneen, 2021; Zhuang et al., 2022; Srivastava et al., 2024). Compiler and kernel work addresses transformations, formal equivalence, and bug detection (Chen et al., 2018; Jia et al., 2019; Yang et al., 2021; Pham et al., 2019; Dubey et al., 2025; Qiu et al., 2026); attention studies examine numerical differences and low-precision failure (Golden et al., 2024; Qiu & Yao, 2026). Golden et al. (2024) contextualize Flash Attention's numerical deviation using Wasserstein-based weight-impact proxies. Our empirical comparison centers on frozen substitution decisions, repeated same-state audits, and a confirmed non-exact acceptance with measured full-step speedup. These decision-level tests complement analyses of numerical deviation and its training effects.

Training dynamics and early prediction. Prior methods predict configuration-, run-, or fault-level instability (Kim et al., 2026; Qiu et al., 2025; Huang et al., 2026), while optimizer and representation analyses study training dynamics (Kingma & Ba, 2015; Reddi et al., 2018; Loshchilov & Hutter, 2019; Raghu et al., 2017; Kornblith et al., 2019). RKSP uses an initialization-time spectral score to identify risky training configurations (Kim et al., 2026). Our decision unit is a reference–candidate pair under a declared protocol, including cases in which both runs continue learning; the target is preservation of the reference requirements. Appendix Table 12 contrasts the decision units.

8 Discussion and Conclusion

TPE connects numerical implementation validation to the behavior required during training. In the fresh cohort, its frozen step-50 screen matches every final contract outcome and resolves both types of error made by the fixed-tolerance repeated-local policy. This supports paired-path screening as a practical complement to numerical qualification. Independent confirmation and full-step timing show that an accepted substitution retains a practical efficiency benefit.

The fresh study establishes repeatability on five new seed IDs across the declared conditions; its E2 subset contains seven conditions whose final labels are uniform within each condition. Transfer to unseen implementations, mixed-outcome conditions, and additional non-exact acceptances is the next test of screening generality. The larger-model studies establish rejection under their own protocols, while the confirmed full-step efficiency result concerns the base model. Tolerances express application requirements, and the 50k profile shows why horizon and checkpoint aggregation belong in the decision. Together, the studies support a concrete reference-preservation workflow and identify the workload coordinates that require further validation.

AI Use Statement

Generative AI tools were used to draft sections of the paper and to aid and polish its writing. The authors take responsibility for the final content.

REPRODUCIBILITY STATEMENT

Appendix Table 34 and Appendix A specify protocols, seed roles, aggregation, local-audit definitions, and timing. The artifact maps identify the underlying study records and distinguish manuscript sources from the separate experiment archives, including the E1–E3 run-level records and common probes (Appendix A.5).

REFERENCES

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. URL https://arxiv.org/abs/1607.06450.

Tianqi Chen, Thierry Moreau, Ziheng Jiang, Lianmin Zheng, Eddie Yan, Meghan Cowan, Haichen Shen, Leyuan Wang, Yuwei Hu, Luis Ceze, Carlos Guestrin, and Arvind Krishnamurthy. TVM: An automated end-to-end optimizing compiler for deep learning. In *13th USENIX Symposium on Operating Systems Design and Implementation*, 2018. URL https://www.usenix.org/conference/osdi18/presentation/chen.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Re. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/67d57c32e20fd0a7a302cb81d36e40d5-Abstract-Conference.html.

Kshitij Dubey, Benjamin Driscoll, Anjiang Wei, Neeraj Kayal, Rahul Sharma, and Alex Aiken. Equivalence checking of ML GPU kernels. *arXiv preprint arXiv:2511.12638*, 2025. URL https://arxiv.org/abs/2511.12638.

Ronen Eldan and Yuanzhi Li. TinyStories: How small can language models be and still speak coherent english? *arXiv preprint arXiv:2305.07759*, 2023. URL https://arxiv.org/abs/2305.07759.

David Goldberg. What every computer scientist should know about floating-point arithmetic. *ACM Computing Surveys*, 23(1), 1991. doi: 10.1145/103162.103163. URL https://doi.org/10.1145/103162.103163.

Alicia Golden, Samuel Hsia, Fei Sun, Bilge Acun, Basil Hosmer, Yejin Lee, Zachary DeVito, Jeff Johnson, Gu-Yeon Wei, David Brooks, and Carole-Jean Wu. Is flash attention stable? *arXiv preprint arXiv:2405.02803*, 2024. URL https://arxiv.org/abs/2405.02803.

Suyog Gupta, Ankur Agrawal, Kailash Gopalakrishnan, and Pritish Narayanan. Deep learning with limited numerical precision. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, 2015. URL https://proceedings.mlr.press/v37/gupta15.html.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. URL https://arxiv.org/abs/1709.06560.

Ruixuan Huang, Hantao Huang, Yifan Huang, Ansheng You, Zhenxing Zhang, and Shuai Wang. Mechanism-driven monitors for preemptive detection of LLM training instability. *arXiv preprint arXiv:2606.28116*, 2026. URL https://arxiv.org/abs/2606.28116.

Zhihao Jia, Oded Padon, James Thomas, Todd Warszawski, Matei Zaharia, and Alex Aiken. TASO: Optimizing deep learning computation with automatic generation of graph substitutions. In *Proceedings of the 27th ACM Symposium on Operating Systems Principles*, 2019. doi: 10.1145/3341301.3359630. URL https://doi.org/10.1145/3341301.3359630.

Bum Jun Kim, Shohei Taniguchi, Makoto Kawano, Yusuke Iwasawa, and Yutaka Matsuo. Residual Koopman spectral profiling for predicting and preventing transformer training instability. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 3036–3060. PMLR, 2026. URL https://proceedings.mlr.press/v337/kim26d.html.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey E. Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *5th International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Byj72udxe.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory F. Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. Mixed precision training. In *6th International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1gs9JgRZ.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Koepf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html.

Hung Viet Pham, Thibaud Lutellier, Weizhen Qi, and Lin Tan. CRADLE: Cross-backend validation to detect and localize bugs in deep learning libraries. In *2019 IEEE/ACM 41st International Conference on Software Engineering*, 2019. doi: 10.1109/ICSE.2019.00107. URL https://doi.org/10.1109/ICSE.2019.00107.

Haiquan Qiu and Quanming Yao. Why low-precision transformer training fails: An analysis on flash attention. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=0jHyEKHDyx.

Haiquan Qiu, You Wu, Yingjie Tan, Yaqing Wang, and Quanming Yao. Spectral alignment as predictor of loss explosion in neural network training. *arXiv preprint arXiv:2510.04202*, 2025. URL https://arxiv.org/abs/2510.04202.

Yuxin Qiu, Jiyuan Wang, Ronak Badhe, Ben Limpanukorn, Miryung Kim, and Qian Zhang. Finding compiler-platform interaction bugs in deep learning pipelines via cross-layer constraints. *arXiv preprint arXiv:2606.18421*, 2026. URL https://arxiv.org/abs/2606.18421.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html.

Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *6th International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=ryQu7f-RZ.

Megha Srivastava, Simran Arora, and Dan Boneh. Optimistic verifiable training by controlling hardware nondeterminism. *arXiv preprint arXiv:2403.09603*, 2024. URL https://arxiv.org/abs/2403.09603.

Cecilia Summers and Michael J. Dinneen. Nondeterminism and instability in neural network optimization. *arXiv preprint arXiv:2103.04514*, 2021. URL https://arxiv.org/abs/2103.04514.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1706.03762.

Yichen Yang, Phitchaya Mangpo Phothilimthana, Yisu Wang, Max Willsey, Sudip Roy, and Jacques Pienaar. Equality saturation for tensor graph superoptimization. In *Proceedings of Machine Learning and Systems*, volume 3, 2021. URL https://proceedings.mlsys.org/paper_files/paper/2021/hash/cc427d934a7f6c0663e5923f49eba531-Abstract.html.

Donglin Zhuang, Xingyao Zhang, Shuaiwen Song, and Sara Hooker. Randomness in neural network training: Characterizing the impact of tooling. In *Proceedings of Machine Learning and Systems*, volume 4, pp. 316–336, 2022. URL https://proceedings.mlsys.org/paper_files/paper/2022/hash/427e0e886ebf87538afdf0badb805b7f-Abstract.html.

A Fresh Matched Confirmation, Local-Audit Baselines, and Timing

This section reports the added E1–E3 study. E1 supplies condition-matched reference calibration and fresh paired confirmation. E2 records repeated local audits within the E1 trajectories. E3 is a separate timing experiment. Tables report the frozen study aggregates; Section A.5 maps the supporting experiment records.

A.1 Frozen Design and Counting Units

Table 5: Frozen E1–E3 configuration. Reference-calibration, confirmation, and timing roles use disjoint seed IDs.

Coordinate	Setting
Hardware / system	RTX 4090, 24 GB; Ubuntu 24.04 / WSL
Runtime	Python 3.12.3; PyTorch 2.6.0+cu124; CUDA 12.4; cuDNN 90100
Model	477,696 parameters; width 128; 2 layers; 4 heads; vocabulary 256
Data	Frozen TinyStories and WikiText2 train/validation text slices
Sequence / batch	128 tokens / 16 sequences
Training	5,000 optimizer steps; BF16 autocast; FP32 parameters and native optimizer state
AdamW	Betas $(0.9, 0.999)$; epsilon 10^{-8}; weight decay 0.01
Learning rate	3×10^{-3} default; 10^{-4} low LR; constant schedule
Other	No gradient clipping or loss scaling; initialization and data order matched within each pair
Calibration seeds	34001, 34002, 34003, 34004, 34005
Confirmation seeds	34101, 34102, 34103, 34104, 34105
Timing seeds	34201, 34202, 34203, 34204, 34205

The nine conditions are A0/N0/O0/T0 at the default learning rate on each dataset, plus WikiText2 A0 at the lower learning rate. Each condition has five reference calibration trajectories and five reference–candidate confirmation pairs. Thus E1 comprises 45 reference trajectories plus 45 pairs, or 135 single-branch 5,000-step trajectories represented by 90 scheduled tasks. E2 covers the 35 A0/N0/T0 confirmation units. Its three cumulative reporting nodes produce 105 aggregate rows for the same 35 units. Five seed IDs are reused across conditions; counts are condition–seed outcomes. The original thresholds, seeds, and condition list are retained throughout this study.

A.2 E1: Condition-Matched Seed Calibration

For each condition, the five reference calibration runs supply ten pairwise gaps. Their mean is the diagnostic denominator b_m in Equation 10. Both this denominator and the paired confirmation numerator are computed on one fixed common validation probe: four batches of 16 sequences of length 128, or 8,192 token positions. Saved loss, hidden activations, and logits support loss-gap, hidden-disagreement $(1 - \mathrm{CKA})$, symmetric logit-KL, and prediction-disagreement measurements. Calibration and confirmation roles are disjoint. The mean denominator is not a 95% coverage bound, and the ten pairwise comparisons are dependent through their five source runs.

The original binary contract remains Equation 5, evaluated on the separately retained legacy evaluation records. The new common-probe ratios do not relabel those outcomes. Table 6 reports medians across the five confirmation seeds for each condition.

Table 6: **All nine matched E1 conditions.** Reject counts use the original frozen raw rule; continuous ratios use the new common probe and condition-matched reference-seed means. Tiny T0 hidden ratios reflect CKA floating-point rounding.

Dataset / LR	Pair	Reject	D_{loss}	D_{hidden}	D_{logit}	D_{pred}
TinyStories / default	A0	5/5	0.2091	0.8412	0.7095	0.8239
TinyStories / default	N0	5/5	1.142	0.8260	0.7189	0.8080
TinyStories / default	O0	5/5	0.2728	0.8544	0.6978	0.8285
TinyStories / default	T0	0/5	0	6.761×10^{-15}	0	0
WikiText2 / default	A0	5/5	1.697	0.8576	0.7499	0.8241
WikiText2 / default	N0	5/5	0.1452	0.8992	0.7209	0.8122
WikiText2 / default	O0	5/5	0.3940	0.8904	0.7008	0.8030
WikiText2 / default	T0	0/5	0	3.041×10^{-15}	0	0
WikiText2 / low LR	A0	0/5	0.004439	0.0002135	0.001251	0.02988

The study reproduces all default A0/N0/O0 rejections, all exact-control passes, and all lower-learning-rate A0 passes. For the accepted A0 condition, every reported median functional ratio is below 0.03. Default O0's loss and prediction ratios are below one on both datasets. This revises the earlier O0-specific seed-band interpretation while preserving its reproduced raw-contract outcomes. The matched E1 denominators are used only with their matching fresh numerators; they are not applied to earlier hardware cohorts or to the 50,000-step trajectories.

A.3 E2: Repeated Re-Anchored Local Audits

At steps 0/50/100/200, the local audit places both implementations at the current reference-branch parameter state and applies the unchanged forward/backward gate to four fixed batches. It does not compare the two already-separated parameter states. Each cumulative local decision rejects after any observed batch failure. The independent paired-path screen instead measures gradient disagreement between the branches at their own step-50 states and applies the previously frozen threshold $1.8477439880371094 \times 10^{-5}$. The step-50 decision is retained at subsequent reporting nodes.

Table 1 reports the complete confusion matrices. The 20 positive units are default A0/N0 on the two datasets; the 15 negative units are default T0 on the two datasets and lower-learning-rate WikiText2 A0. Local audits are verified to leave the live training state and RNG unchanged. O0 is excluded because its substitution occurs in the optimizer, for which identical forward/backward outputs do not provide a meaningful local optimizer check.

Exact-control and non-exact strata. Table 7 separates the role of exact controls from non-exact substitutions. Counts follow directly from Table 1 and the condition-level outcomes: repeated local checks identify all ten default A0 rejections by step 50, identify seven of ten default N0 rejections by step 200, and falsely reject all five lower-learning-rate A0 units by that node. The remaining ten passes are T0. These are breakdowns of the existing 35 units, not additional experiments or independent validation cohorts.

Table 7: E2 outcomes separated by substitution type. Entries count decisions agreeing with the final raw contract; local decisions accumulate checks through step 200.

Stratum	Units	Repeated local / 200	Paired path / 50
Default A0 (non-exact, reject)	10	10/10	10/10
Default N0 (non-exact, reject)	10	7/10	10/10
Low-LR A0 (non-exact, pass)	5	0/5	5/5
T0 (exact, pass)	10	10/10	10/10
All non-exact substitutions	25	17/25	25/25

Traceable false-negative example. For WikiText2 N0, confirmation seed 34101, the reported step-50 gradient cosine is 0.9999702002365346, giving disagreement $2.97997634654 \times 10^{-5}$. The

paired-path screen therefore predicts rejection. All 16 local batch checks through step 200 pass, with unchanged live state and RNG. At step 5,000, parameter relative distance is 0.62298965 and hidden CKA is 0.83469439; both violate the raw contract. The absolute validation-loss gap, 0.00072107, remains below its threshold. This case illustrates a rejection that a longer repeated-local audit misses and the shorter paired path detects.

A.4 E3: Full GPU Step Timing and Validation Cost

E3 separately measures WikiText2 A0 under both learning rates. Each of five ABBA groups runs Math, Flash, Flash, Math, resetting model and optimizer state for every segment. Every segment warms up for 50 steps and then synchronously measures 200 steps. The 40 segments contain 10,000 total training steps, of which 8,000 are measured. Forward, backward, and optimizer update are included; GPU-resident data remove loading and host-to-device transfer from the measured interval.

For group j, the ratio is

$$R_j = \frac{t_{\mathrm{Math,first},j} + t_{\mathrm{Math,last},j}}{t_{\mathrm{Flash,first},j} + t_{\mathrm{Flash,last},j}}. \tag{13}$$

Table 3 reports $\mathrm{median}_j\, R_j$ alongside the separately summarized step times. These summaries need not have the same ratio. The independently measured $1.618\times$ attention-only result remains a different measurement with its original setup.

Table 8: Validation-cost accounting for WikiText2 A0. Full confirmation already contains the prefix and initialization audit. Reuse counts extrapolate short-segment timing to repeated 5,000-step workloads under the same protocol.

Quantity	Default LR	Low LR
Calibration + full confirmation, summed wall time	≈ 402 s	≈ 392 s
Step-50 paired prefix + initial audit, median	0.839 s	0.829 s
Full-step ABBA speedup, median	$1.348\times$	$1.357\times$
Estimated same-protocol reuses for amortization	≈ 50	≈ 51

Calibration means five full reference runs; confirmation means five full reference–candidate pairs. Prefix and audit costs are already included in the full accounting and are not added again. The reuse estimate is a workload-specific extrapolation of measured step savings, not a measured deployment break-even across tasks. The timing excludes data-pipeline overhead and concerns this small model; the full validation cost is material despite the short early prefix.

Timing-record recovery. All 40 timing segments completed before a final comparison reported failure: AdamW betas were represented as a Python tuple before JSON serialization and a list afterward. The archived custody check attributes the discrepancy to those 40 type representations, retains all original numeric and segment data, and applies the frozen summarizer in a CPU-only recovery. The documented valid entry point is `benchmark/results/current.json`; the original failure and recovery records remain part of the experiment archive.

A.5 Evidence Provenance and Archive Map

The E1–E3 experiment archive is separate from the manuscript source package. It organizes the frozen protocol, source/runtime hashes, public-probe outputs, input snapshots, and run records; model and optimizer checkpoints are excluded. Table 9 maps its documented evidence paths. The manuscript package contains the aggregate tables, while the full archive supports run-level inspection and re-aggregation.

Table 9: Documented paths in the separate E1–E3 experiment archive. These paths locate evidence; they do not imply that those files are included in the manuscript source package.

Evidence	Archive location
Frozen design / source binding	`frozen/protocol.json`; `frozen/execution_manifest.json`
Training / audit implementation	`runner.py`; `frozen_source/`
Run-level outcomes and probes	`runs/<condition>__<role>__seed<id>/` with `summary.json`, `metrics.jsonl`, and `probe.pt`
Completed E1/E2 aggregates	`reports/latest.json` points to `reports/complete/<hash>/`
E1 tables	`e1.csv`; `baselines.csv`; `calibration_pairs.csv` in the aggregate directory
E2 tables	`e2.csv`; `strata.csv`; separate optimizer diagnostic `e2_o0.csv`
Example seed record	`quick_results/E2_example_seed34101.json`
E3 timing source	`benchmark/run_e3.py`; `cuda_timing.py`; `e3_design.py`
Raw timing segments	`benchmark/results/e3/segment_000.json` through `segment_039.json`
Timing summary and recovery	`benchmark/results/current.json`; `e3_recovered/summary.json`; `e3_recovered/recovery.json`
Cost accounting	`analysis/cost_accounting.py`; `analysis/cost_accounting_report.md`
Integrity / handoff	`BUNDLE_MANIFEST.json`; `verify_bundle.py`; `HANDOFF_STATUS.json`

Reading aggregate tables and verifying the archive manifest require no GPU. Re-aggregating saved common probes requires PyTorch and can use the CPU. The E2 aggregate has 105 rows because each of the 35 units appears at three reporting nodes; E1 has 45 paired confirmation rows, with reference calibration represented separately. This separation of cohorts, evaluation records, and counting units underlies the comparisons in the main text.

B Screening and Diagnostic Displays

This appendix retains the earlier screening displays, confirmations, mechanism studies, and robustness analyses. These historical cohorts are separate from the fresh E1–E3 evaluation in Appendix A.

B.1 Evaluation Workflow and Earlier Main-Text Displays

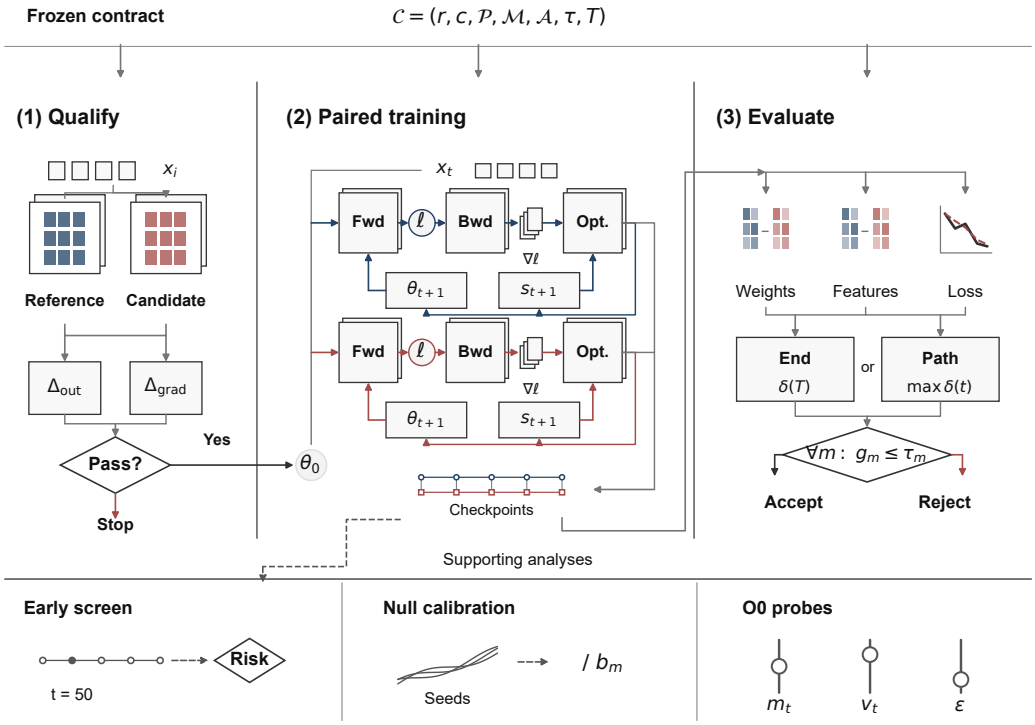

Figure 3: **TPE evaluation under a frozen substitution contract.** **(1)** Apply the declared output and gradient qualification gate. **(2)** Train the pair from matched initialization and data order; each branch performs its own forward, backward, and optimizer updates and retains separate parameters and optimizer states. **(3)** Compare weights, features, and validation loss at declared checkpoints. Use the contract's terminal or path-uniform aggregation and accept only when every required gap meets its tolerance. The lower panels show supporting analyses: early pairwise screening, matched-seed diagnostic normalization, and O0 state/epsilon interventions. These analyses do not replace the raw acceptance rule. Icons and traces are schematic.

Table 10: Local qualification followed by training-time rejection. All cells use the 5,000-step, $n = 5$ real-data protocol. The real A0 kernel pair and all three replication pairs diverge under default AdamW, while T0 and SGD controls remain stable.

Dataset	Pair	n	Default div	10^{-2} div	SGD div	Default param	Default hidden CKA
TinyStories	A0	5	5/5	2/5	0/5	0.5949	0.8552
TinyStories	A1	5	5/5	3/5	0/5	0.6011	0.8551
TinyStories	N0	5	5/5	4/5	0/5	0.5999	0.8526
TinyStories	O0	5	5/5	3/5	0/5	0.5971	0.8521
TinyStories	T0	5	0/5	0/5	0/5	0.0000	1.0000
WikiText2	A0	5	5/5	0/5	0/5	0.6238	0.8466
WikiText2	A1	5	5/5	0/5	0/5	0.6264	0.8438
WikiText2	N0	5	5/5	0/5	0/5	0.6240	0.8442
WikiText2	O0	5	5/5	0/5	0/5	0.6174	0.8438
WikiText2	T0	5	0/5	0/5	0/5	0.0000	1.0000

Table 11: A0 acceptance boundary on WikiText2. The frozen step-50 rule, local gate, BF16 AdamW, pair, data, and 5,000-step horizon are shared. Kernel speed is measured independently on the indicated hardware stack.

LR	Screen risk	Speedup	Stable	Param rel	Hidden CKA	Loss gap	Val. reduction
3×10^{-3}	2.545×10^{-5}	1.465×	0/5	0.6238	0.8466	0.002566	–
10^{-4}	1.490×10^{-6}	1.618×	**5/5**	0.0002	1.0000	2.033×10^{-4}	3.513

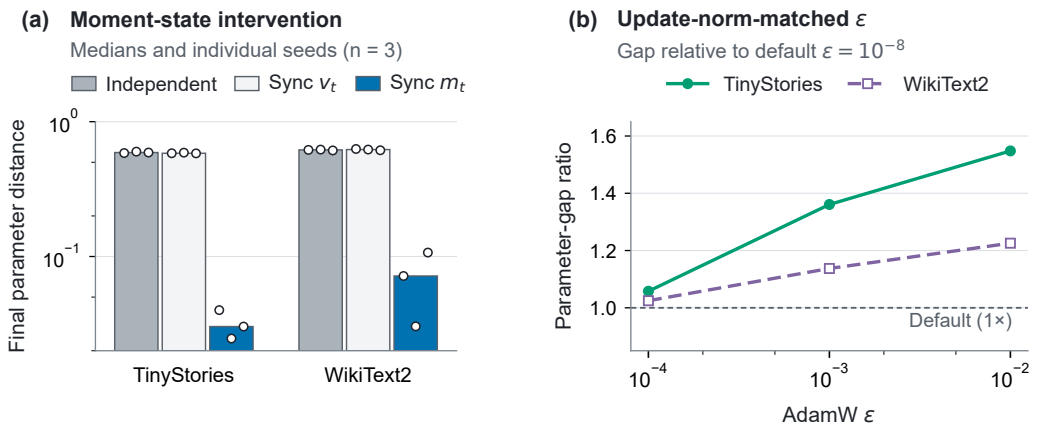

Figure 4: O0 optimizer-state interventions and update-matched epsilon controls. **(a)** Synchronizing the first moment m_t, but not the second moment v_t, suppresses final parameter separation. Bars are three-seed dataset medians; circles show the individual seed outcomes. **(b)** After matching update norms, increasing epsilon does not reduce the parameter gap relative to default AdamW at $\epsilon = 10^{-8}$; the dashed horizontal line marks a ratio of one.

Table 12: Comparison with closely related work. Pair conditioning denotes an explicit reference-relative decision.

Work	Decision unit	Pair-conditioned?	Target	Stage	Output
RKSP Kim et al. (2026)	Configuration	No	Training divergence	Initialization	Predict/intervene
Spectral alignment Qiu et al. (2025)	Run	No	Loss explosion	Training	Warn
Mechanism monitors Huang et al. (2026)	Run/fault	No	Specific instability	Training	Warn
VOLTA Dubey et al. (2025)	Kernel pair	Yes	Formal equivalence	Pre-training	Verify
XCheck Qiu et al. (2026)	Compiler/platform case	Varies	Bugs/classes	Testing	Detect
TPE	Pair + protocol	Yes	Learning-time substitution	Early training	Accept/reject

Table 13: Native O0 state interventions, reported as three-seed medians. Synchronizing the first moment, not the second moment alone, suppresses separation while retaining the fused/unfused optimizer kernels.

Dataset	Condition	Param rel	Hidden CKA	Loss gap	Pred disag.
TinyStories	independent	0.5906	0.8562	0.00345	0.1680
TinyStories	sync $\hat{v}$	0.5833	0.8489	0.00463	0.1758
TinyStories	sync $\hat{m}$	0.0302	0.9992	0.00120	0.0132
WikiText2	independent	0.6179	0.8468	0.00792	0.1953
WikiText2	sync $\hat{v}$	0.6213	0.8454	0.01147	0.2017
WikiText2	sync $\hat{m}$	0.0717	0.9992	0.00031	0.0166

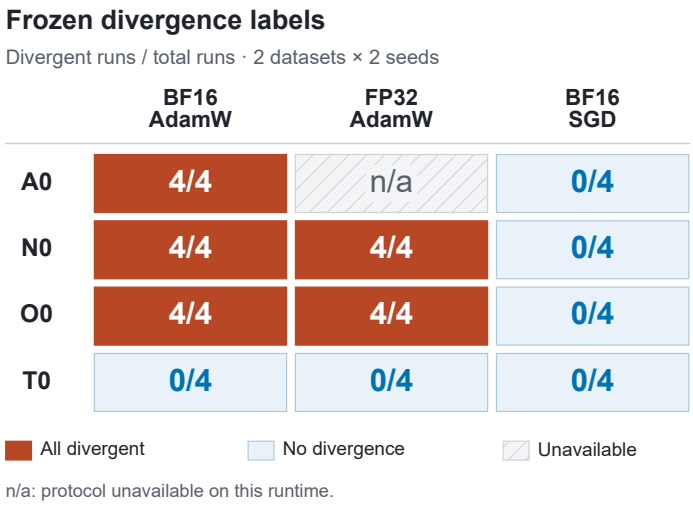

Figure 5: Pre-specified divergence-label matrix from the earlier default screen. Counts aggregate two datasets and two seeds. O0 diverges under default AdamW despite zero local output and gradient differences, while SGD and T0 controls remain stable. "n/a" denotes protocol cells unavailable on this runtime.

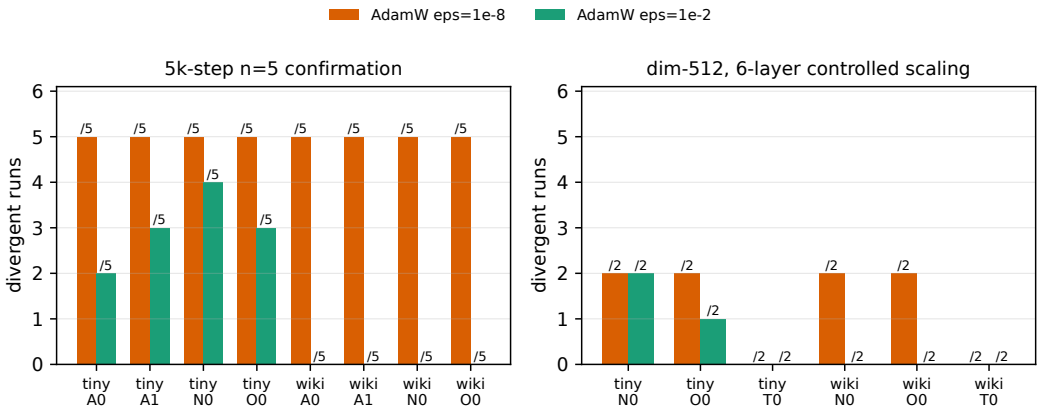

Figure 6: Persistence across seeds and model scale under controlled protocols. The left panel includes A0/A1/N0/O0 five-seed confirmations; the right panel is a scoped dimension-512, six-layer check complementing the larger-model replications.

Table 14: Frozen reference-conditioned decision screen on WikiText2. Speed is forced-backend BF16 forward/backward throughput relative to Math-SDPA. All candidates pass the pre-specified initialization-state local gate. The step-50 risk threshold (1.848×10^{-5}) is fit on existing TinyStories runs and is not refit for any listed decision.

Pair	Candidate	Protocol	Speedup	95% lower	Step-50 risk	Long result
A0	Flash	3×10^{-3}	$1.465\times$	$1.386\times$	2.545×10^{-5}	5/5 div.
A2	Efficient	3×10^{-3}	$1.427\times$	$1.343\times$	4.023×10^{-5}	5/5 div.
A3	cuDNN	3×10^{-3}	$1.529\times$	$1.349\times$	5.764×10^{-5}	rejected
A0	Flash	10^{-4}	$1.618\times$	$1.561\times$	1.490×10^{-6}	**5/5 stable**

Table 15: Functional separation for the five-seed A2 long validation on WikiText2. Values are run medians.

n	Param rel	Hidden CKA	Loss gap	Logit sym. KL	Pred. disagree
5	0.6274	0.8450	0.0072	0.1650	0.1963

Table 16: Timing and memory stack for the SDPA decision probe. Measurements use an RTX 6000D, PyTorch 2.8.0+cu128, CUDA 12.8, and cuDNN 9.10.2; each forced-backend BF16 forward/backward microbenchmark uses 50 warmups, 100 repeats, and shape $[16, 4, 128, 32]$. The lower bound is the 2.5th percentile of 2,000 independent latency-resampling draws. Memory is peak CUDA allocator usage within the measured backend loop.

Backend	Median ms	P95 ms	Speedup	95% lower	Peak MiB
Math	1.103	1.213	$1.000\times$	$0.943\times$	45.75
Flash	0.753	0.838	$1.465\times$	$1.386\times$	31.78
Efficient	0.773	0.875	$1.427\times$	$1.343\times$	31.78
cuDNN	0.721	0.894	$1.529\times$	$1.349\times$	31.78

Table 17: Independent RTX 4090 Math/Flash timing used for the accepted A0 decision and large-model context. PyTorch is 2.6.0+cu124 with CUDA 12.4. Each forced-backend BF16 forward/backward probe uses 100 warmups and 300 repeats. The accepted base-model decision uses the $[16, 4, 128, 32]$ row. The 355M microbatch row is a separate shape-matched kernel probe; across four AB/BA 355M training observations, the median end-to-end trajectory speedup is $1.358\times$.

Shape	Math ms	Flash ms	Speedup	95% lower
$[16, 4, 128, 32]$	0.771	0.477	$1.618\times$	$1.561\times$
$[1, 20, 128, 64]$	0.795	0.374	$2.127\times$	$1.781\times$

Table 18: Re-anchored local audit on WikiText2. At each listed training step, both implementations are reconstructed from the same reference-branch parameter state and checked on four fixed batches over five seeds. "Later" aggregates steps 50, 100, and 200. The audit is stronger than the initialization-state qualification gate used by the prospective screen.

Pair	Init pass	Later pass	Worst output rel	Worst grad rel	Worst grad cosine
A0	20/20	0/60	0.00685	0.00407	0.999996
A1	20/20	0/60	0.02339	0.01476	0.999894
N0	20/20	60/60	0.00671	0.00191	0.999999
T0	20/20	60/60	0.00000	0.00000	1.000000

The re-anchored audit shows how local agreement changes along training. N0 and T0 pass at every tested state. A0 retains sub-0.7% output disagreement and near-unit gradient cosine, but its relative-

gradient statistic exceeds the initialization threshold after training begins; A1 also has a step-200 worst case. The main results use initialization-state qualification and do not assume that this gate continues to pass later. Repeating local checks at trained reference states adds path-dependent information that a single initialization check cannot provide.

Table 19: Preregistered Phase24 L8/L4 boundary matrix. "Original" denotes the frozen TPE rule; "augmented" additionally audits LayerNorm affine parameters and optimizer state. All 40 paired runs pass the non-static defense. The preceding 3×10^{-3} stress experiment gives 0/5 original-rule stable runs for every listed pair/dataset.

Dataset	Pair	LR	Original	Augmented	Param rel	Hidden CKA	Loss gap	Val. reduction
WikiText2	L8	3×10^{-4}	4/5	0/5	0.0033	0.9992	3.398×10^{-4}	0.703
WikiText2	L8	10^{-4}	5/5	4/5	0.0002	1.0000	1.761×10^{-4}	0.612
WikiText2	L4	3×10^{-4}	0/5	0/5	0.0033	0.9991	1.793×10^{-3}	0.703
WikiText2	L4	10^{-4}	5/5	4/5	0.0002	1.0000	1.484×10^{-4}	0.612
TinyStories	L8	3×10^{-4}	2/5	0/5	0.0041	0.9987	7.722×10^{-4}	0.769
TinyStories	L8	10^{-4}	3/5	1/5	0.0004	0.9999	4.654×10^{-4}	0.660
TinyStories	L4	3×10^{-4}	0/5	0/5	0.0046	0.9987	3.207×10^{-4}	0.769
TinyStories	L4	10^{-4}	2/5	1/5	0.0003	0.9999	6.609×10^{-4}	0.660

The boundary experiments run in a dedicated environment on an NVIDIA A100-SXM4-80GB with PyTorch 2.6.0+cu124. The artifact package contains the preregistration, executable snapshots, per-run telemetry, and complete attempt ledger. The reported counts include all attempts, including incomplete or failed runs.

Table 20: Controlled dimension-512, six-layer scaling evaluation. Each cell uses two seeds and 3,000 steps; the study complements the larger-model replications.

Dataset	Pair	Default div	10^{-2} div	Param rel default $\to 10^{-2}$	Hidden CKA default $\to 10^{-2}$
TinyStories	N0	2/2	2/2	1.1350 $\to$ 0.0021	0.1648 $\to$ 0.9963
TinyStories	O0	2/2	1/2	1.0985 $\to$ 0.0023	0.4151 $\to$ 0.9969
TinyStories	T0	0/2	0/2	0.0000 $\to$ 0.0000	1.0000 $\to$ 1.0000
WikiText2	N0	2/2	0/2	1.1040 $\to$ 0.0004	0.2595 $\to$ 0.9999
WikiText2	O0	2/2	0/2	1.1031 $\to$ 0.0003	0.3245 $\to$ 0.9999
WikiText2	T0	0/2	0/2	0.0000 $\to$ 0.0000	1.0000 $\to$ 1.0000

Table 21: Conditional-acceptance screening under unmatched 10^{-2}. The Phase18 step-50 rule and speed probe are reused without refitting. Both candidates remain above the frozen risk threshold, so the pre-registered policy launches no five-seed acceptance run.

Pair	Candidate	Speedup	Step-50 risk	Promote
A2	Efficient	1.427×	1.931×10^{-5}	no
A3	cuDNN	1.529×	2.068×10^{-5}	no

Table 22: Robust denominator summaries for paired default-to-10^{-2} comparisons. Raw log-scale quantiles remain in the telemetry artifact.

Dataset	Pair	n	Param rel default $\to 10^{-2}$	Denom rel default $\to 10^{-2}$	Extreme p95 runs
TinyStories	A1	3	0.5988 $\to$ 0.0036	0.1970 $\to$ 0.0037	0/3 $\to$ 0/3
TinyStories	N0	3	0.5997 $\to$ 0.0032	0.1722 $\to$ 0.0036	0/3 $\to$ 0/3
TinyStories	O0	3	0.5998 $\to$ 0.0033	0.2581 $\to$ 0.0034	0/3 $\to$ 0/3
WikiText2	A1	3	0.6238 $\to$ 0.0003	0.2422 $\to$ 0.0002	0/3 $\to$ 0/3
WikiText2	N0	3	0.6207 $\to$ 0.0003	0.2165 $\to$ 0.0002	0/3 $\to$ 0/3
WikiText2	O0	3	0.6184 $\to$ 0.0003	0.2374 $\to$ 0.0002	2/3 $\to$ 1/3

Table 23: Absolute validation performance of O0 under the unmatched epsilon intervention. 10^{-2} reduces implementation disagreement while increasing validation loss and implied perplexity.

Dataset	Eps	Div	Val loss	PPL	Loss gap
TinyStories	10^{-8}	5/5	1.0659	2.9034	0.0038
TinyStories	10^{-2}	3/5	1.5145	4.5470	0.0012
WikiText2	10^{-8}	5/5	1.3930	4.0269	0.0043
WikiText2	10^{-2}	0/5	2.1914	8.9477	0.0002

Table 24: Update-norm-matched O0 epsilon ladder. Norm ratios are medians over six runs; dataset parameter gaps are three-seed medians. Default-epsilon gaps are 0.5906 on TinyStories and 0.6179 on WikiText2.

Epsilon	Matched	Update-norm ratio	TinyStories param	WikiText2 param
10^{-4}	6/6	0.926	0.6346	0.6364
10^{-3}	6/6	0.946	0.8037	0.7027
10^{-2}	6/6	0.955	0.9142	0.7502

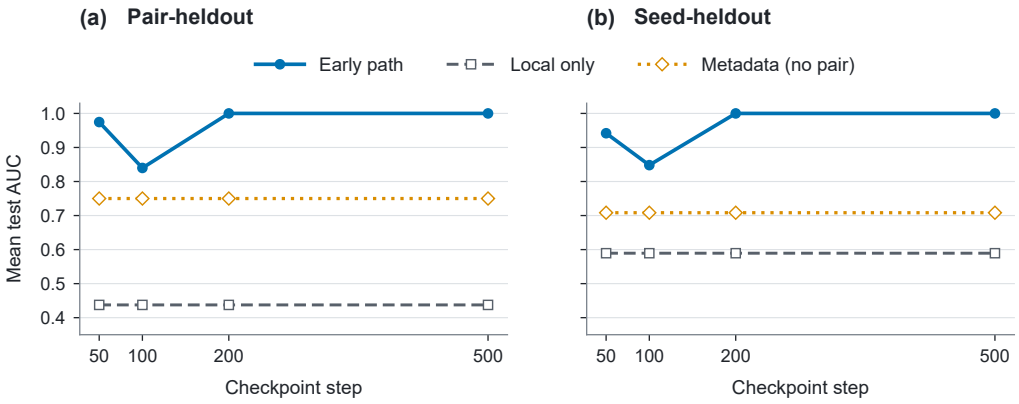

Figure 7: In-matrix screening across early checkpoints. The stricter fixed-optimizer, cross-domain, and cross-scale tests are reported in Table 2.

Paired early-minus-local AUC differences are 0.151 [−0.100, 0.389] for TinyStories, 0.267 [0.103, 0.468] for WikiText2, and 0.234 [0.078, 0.416] for dimension-512 transfer. Extending the prefix from step 50 to steps 50–200 raises cross-scale AUC to 0.972. All feature choices use training rows only.

Table 25: Statistical inference and frozen-threshold results for early-path transfer. All feature directions are positive; thresholds are fit on training rows only. Sens., spec., prec., and NPV are evaluated once on the held-out rows.

Held out	Feature	Threshold	Train/test	Perm. p	Sens.	Spec.	Prec.	NPV
TinyStories	Hidden disagreement	7.153×10^{-7}	16/16	0.271	0.667	0.857	0.857	0.667
WikiText2	Gradient disagreement	1.848×10^{-5}	16/16	0.059	1.000	0.700	0.667	1.000
Dim-512	Logit sym. KL	1.759×10^{-5}	32/24	0.031	1.000	0.615	0.688	1.000

C PREREGISTERED A0 LONG-HORIZON EXTENSION

We extend the lower-learning-rate A0 confirmation from the original terminal $T = 5{,}000$ contract to a separately preregistered path-uniform contract through 50,000 optimizer steps. The implementation pair, five fresh seeds, raw thresholds, and evaluation procedure remain unchanged. Table 26 reports the declared checkpoints. Each seed crosses the frozen raw threshold at least once between steps 7,000 and 17,000, violating the requirement that every seed pass at every checkpoint. All five nevertheless pass at step 5,000 and again at the terminal 50,000-step checkpoint. The extension therefore preserves the original terminal decision while showing why horizon and path aggregation must be specified separately.

Figure 8 places the five-seed extension alongside the separate three-seed learning-rate boundary screen.

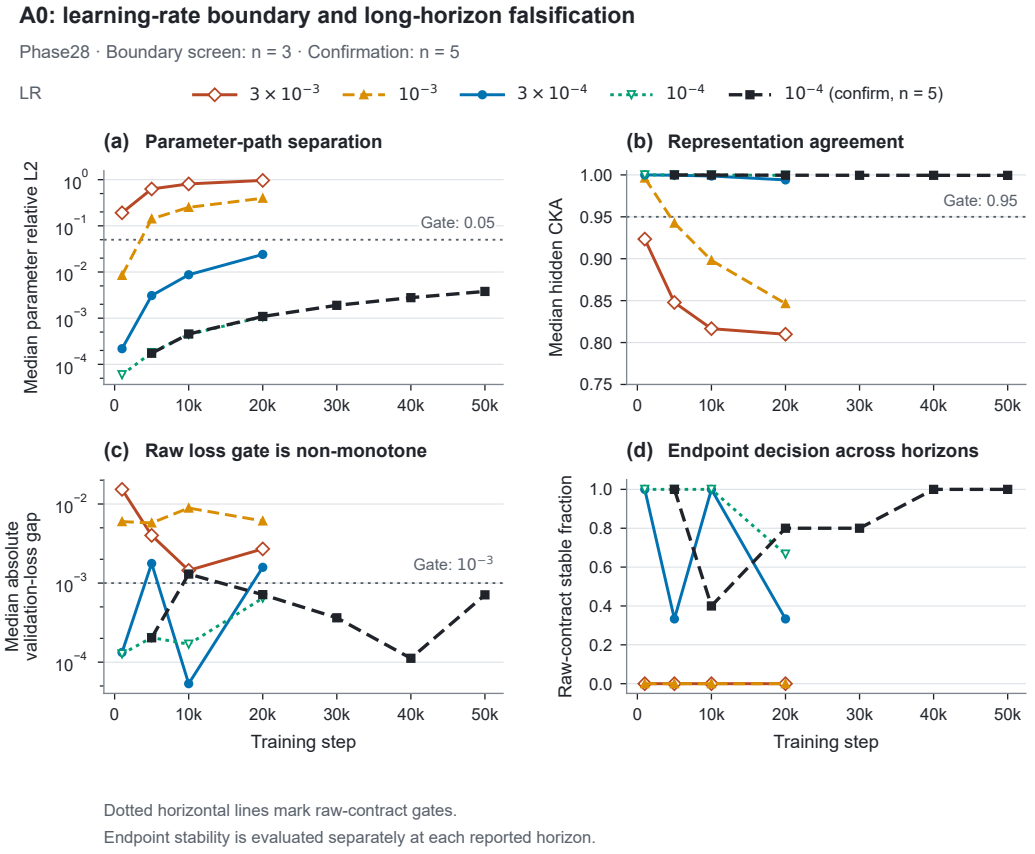

Figure 8: Learning-rate and horizon dependence for the public A0 Math-SDPA/Flash-SDPA pair. **(a–c)** Median parameter distance, hidden CKA, and absolute validation-loss gap. Colored curves show the three-seed boundary screen through 20,000 steps; the black dashed curve shows the five-seed confirmation through 50,000 steps. Dotted horizontal lines mark the frozen raw thresholds. **(d)** Fraction of seeds satisfying the raw contract at each reported endpoint. Endpoint stability is distinct from path-uniform acceptance.

Table 26: Preregistered 50,000-step extension for A0 under BF16 AdamW at learning rate 10^{-4} on an RTX 4090. Stable counts use the unchanged frozen raw rule; remaining entries are five-seed medians at each declared report checkpoint. The terminal checkpoints pass, but the preregistered path-uniform contract fails because of intermediate crossings.

Step	Stable	Param rel	Hidden CKA	Loss gap
5,000	5/5	0.000174	0.999993	2.033×10^{-4}
10,000	2/5	0.000454	0.999820	1.301×10^{-3}
20,000	4/5	0.001090	0.999613	7.142×10^{-4}
30,000	4/5	0.001899	0.999539	3.640×10^{-4}
40,000	5/5	0.002784	0.999528	1.120×10^{-4}
50,000	5/5	0.003795	0.999449	7.102×10^{-4}

Table 27: Tolerance profile for the preregistered A0 50,000-step extension. Multiplier α scales the parameter and hidden-disagreement tolerances from 0.05 and the validation-loss-gap tolerance from 0.001. "Report" uses the six preregistered report checkpoints; "dense" additionally audits every recorded 1,000-step checkpoint from 5k to 50k. All observed crossings are loss-gap-only.

α	Param max	CKA min	Loss max	Report pass	Dense pass	50k pass
0.5	0.025	0.975	0.0005	0/5	0/5	1/5
1.0	0.050	0.950	0.0010	0/5	0/5	5/5
2.0	0.100	0.900	0.0020	4/5	2/5	5/5

At the frozen tolerance, all five seeds fail the path-uniform contract, and every violation is due to validation-loss gap. Doubling the tolerances removes all report-grid crossings for four seeds and all dense-grid crossings for two. One report-grid seed and three dense-grid seeds still cross. These pass-count changes identify a tolerance-sensitive contract boundary; parameter and representation gaps remain within their thresholds.

D Independent V100 Large-Scale Replication

The independent V100 study uses FP16 AdamW at learning rate 3×10^{-4}, effective batch 16, sequence length 128, and seeds 37/43/53 with AB/BA/AB execution order. Each group in Table 28 is aggregated separately from the A100 and RTX 4090 studies. Seed bands are never pooled across hardware.

Table 28: Independent V100 large-model replication. Metrics are three-seed medians at the indicated horizon. "Local" counts frozen local-gate passes and "Div." counts frozen raw-rule divergences.

Scale	Data	Pair	Steps	Local	Div.	Param rel	Hidden CKA	Loss gap
354.92M	TinyStories	O0	500	3/3	3/3	0.0584	0.8895	0.0376
354.92M	WikiText2	O0	500	3/3	3/3	0.0276	0.9692	0.0016
354.92M	TinyStories	L8	500	3/3	3/3	0.0532	0.9506	0.0334
354.92M	WikiText2	L8	500	3/3	3/3	0.0255	0.9789	0.0159
354.92M	TinyStories	T0	500	3/3	0/3	0.0000	1.0000	0.0000
354.92M	WikiText2	T0	500	3/3	0/3	0.0000	1.0000	0.0000
1.003B	WikiText2	O0	500	3/3	0/3	0.0093	0.9772	0.0003
1.003B	WikiText2	O0	1,000	3/3	1/3	0.0148	0.9605	0.0001
1.003B	WikiText2	T0	500	3/3	0/3	0.0000	1.0000	0.0000
1.003B	WikiText2	T0	1,000	3/3	0/3	0.0000	1.0000	0.0000

The 1.003B O0 result depends on the seed: one run reaches hidden CKA 0.819 at step 1,000, while the other two remain within the frozen rule. The result identifies a scale-and-horizon boundary rather than uniform divergence. Separately, 355M N0 is $3/3$ divergent but fails the local-gradient gate in all three runs, with median local gradient relative $\ell_2 = 0.00460$. We therefore exclude it from the positive TPE count.

E ARTIFACT MAP

Table 29: Primary evidence artifacts used to assemble the manuscript.

Role	Artifact
Core evidence and robustness	`phase16_paper_evidence_pack.md`; `phase16_appendix_defense_tables_20260709.md`
Pair definitions and A0 confirmation	`phase17_pair_definition_report.md`; `phase17_a0_realkernel_n5_prediction_report.md`
Fixed-optimizer prediction	`phase17_fixed_adamw_prediction_report.md`; `phase17_fixed_adamw_prediction_manifest.json`
Frozen decision protocol	`phase18_kernel_selection_training_rows_20260711.csv`; `phase18_kernel_selection_rule_20260711.json`
Prospective rejection evidence	`phase18_kernel_selection_long_validation_20260711.md`; `phase20_reproducibility_defense_20260713.md`
Multi-state local audit	`phase19_multistate_local_audit_report_20260713.md`
O0 mechanism evidence	`phase17_evidence_report.md`; `phase19_long_realdata_n5_a0_completed_report.md`
A100 custody and execution	`a100_return/A100_EXECUTION_SUMMARY.md`; `a100_return/COMPLETE_ATTEMPT_LEDGER.csv`
Phase24 boundary and inference	`a100_return/results/phase24/`; `a100_return/results/prediction_pair_seed_corrected/`
4090 accepted A0 decision	`artifacts/4090_a0_positive_decision.json`; `artifacts/A0_ACCEPTANCE_CONFIRMATION_PREREGISTRATION.md`
50k A0 extension and LR boundary	`artifacts/phase28_p0_report.json`; `artifacts/phase28_p0_report.md`
50k tolerance profile	`artifacts/phase28_50k_threshold_sensitivity.json`; `artifacts/phase28_50k_threshold_sensitivity.csv`
355M five-seed replication	`artifacts/phase29_p1_report.json`; `artifacts/4090_350m_summary.json`
4090 timing probes	`artifacts/4090_sdpa_probe_small.json`; `artifacts/4090_sdpa_probe_350m_microbatch.json`
V100 355M/1.003B replication	`artifacts/v100_phase26_summary.json`; custody hashes and raw summaries in the supplementary V100 return
V100 five-seed and horizon extension	`artifacts/phase29_p1_report.json`; `artifacts/v100_phase29_horizon_report.json`
6000D 1.003B A0/T0 replication	`artifacts/phase30_phase33_1b_n5_report.json`; ten run summaries matching `artifacts/scale1b_*_summary.json`
6000D 354.9M low-LR 20k falsification	`artifacts/phase31_355m20k/phase31_report.json`; `artifacts/phase31_355m20k/phase31_runs.csv`; three raw summaries; `artifacts/phase31_355m20k/RESULT_MANIFEST.sha256`
Release index and bundle	`README_ARTIFACT_RELEASE.md`; `TPE_A100_PUBLIC_20260714.tar.gz`

F FULL SEED-NORMALIZED TPE MATRIX

Table 30 retains the historical A0/A1/N0/O0/T0 long-confirmation ratios. The numerator is the original 5,000-step implementation-pair gap; the denominator is the earlier locked reference-seed scale grouped by precision and optimizer. Matching of all dataset, model, runtime, horizon, and probe coordinates is not established for this historical denominator. These values preserve the earlier analysis and are not used as evidence that a pair exceeds fully matched seed variation. The fresh matched comparison is Table 6; binary outcomes retain the raw rule.

Table 30: Historical normalized ratios for the 5,000-step long-confirmation matrix, retained with their original denominator. Fresh matched ratios are reported separately in Table 6.

Dataset	Pair	Opt	Eps	n	D_{loss}	D_{hidden}	D_{logit}	D_{pred}	Param rel
TinyStories	A0	adamw	1e-8	5	3.1767	0.2020	0.8138	1.3537	0.6010
TinyStories	A0	adamw	1e-2	5	0.6965	9.230e-04	0.0074	0.1884	0.0030
TinyStories	A0	sgd	1e-8	5	0.0073	4.521e-07	6.817e-05	5.877e-04	1.517e-06
TinyStories	A1	adamw	1e-8	5	2.9117	0.2007	0.8111	1.3971	0.6011
TinyStories	A1	adamw	1e-2	5	0.6147	0.0011	0.0089	0.1932	0.0032
TinyStories	A1	sgd	1e-8	5	0.0155	6.149e-07	7.094e-05	2.938e-04	1.531e-06
TinyStories	N0	adamw	1e-8	5	3.2254	0.2042	0.8297	1.3971	0.5999
TinyStories	N0	adamw	1e-2	5	0.6925	0.0012	0.0098	0.2286	0.0029
TinyStories	N0	sgd	1e-8	5	0.0101	5.244e-07	6.584e-05	1.959e-04	1.406e-06
TinyStories	O0	adamw	1e-8	5	1.3380	0.2048	0.8343	1.4255	0.5971
TinyStories	O0	adamw	1e-2	5	0.4378	0.0011	0.0088	0.1853	0.0031
TinyStories	O0	sgd	1e-8	5	0.0000	0.0000	0.0000	0.0000	0.0000
TinyStories	T0	adamw	1e-8	5	0.0000	0.0000	0.0000	0.0000	0.0000
TinyStories	T0	adamw	1e-2	5	0.0000	0.0000	0.0000	0.0000	0.0000
TinyStories	T0	sgd	1e-8	5	0.0000	0.0000	0.0000	0.0000	0.0000
WikiText2	A0	adamw	1e-8	5	1.3807	0.2138	0.7818	1.6068	0.6276
WikiText2	A0	adamw	1e-2	5	0.0596	1.066e-05	6.273e-04	0.0647	2.795e-04
WikiText2	A0	sgd	1e-8	5	0.0043	5.606e-07	6.432e-05	8.815e-04	1.509e-06
WikiText2	A1	adamw	1e-8	5	1.7231	0.2163	0.7999	1.6147	0.6264
WikiText2	A1	adamw	1e-2	5	0.0344	1.766e-05	6.758e-04	0.0639	3.196e-04
WikiText2	A1	sgd	1e-8	5	0.0064	3.255e-07	6.773e-05	0.0014	1.567e-06
WikiText2	N0	adamw	1e-8	5	1.9416	0.2158	0.7817	1.5784	0.6240
WikiText2	N0	adamw	1e-2	5	0.0701	2.037e-05	6.781e-04	0.0686	3.084e-04
WikiText2	N0	sgd	1e-8	5	0.0060	1.085e-07	5.806e-05	0.0013	1.433e-06
WikiText2	O0	adamw	1e-8	5	1.5263	0.2163	0.7696	1.5461	0.6174
WikiText2	O0	adamw	1e-2	5	0.0859	1.880e-05	7.117e-04	0.0710	3.127e-04
WikiText2	O0	sgd	1e-8	5	0.0000	0.0000	0.0000	0.0000	0.0000
WikiText2	T0	adamw	1e-8	5	0.0000	0.0000	0.0000	0.0000	0.0000
WikiText2	T0	adamw	1e-2	5	0.0000	0.0000	0.0000	0.0000	0.0000
WikiText2	T0	sgd	1e-8	5	0.0000	0.0000	0.0000	0.0000	0.0000

F.1 HISTORICAL SCREEN DIAGNOSTICS

Historical normalization and the matched update. Table 31 preserves the earlier screen's ratios against its historical denominator. Full matching of that denominator to every dataset, model, runtime, horizon, and evaluation probe is not established. Its O0 loss and prediction ratios above one therefore do not establish an exceedance of matched reference-seed variation. The fresh matched study in Table 6 supplies the relevant comparison: default O0 has both ratios below one on each dataset, while retaining all raw-contract rejections. Table 30 similarly retains the earlier long-confirmation ratios for provenance; the new denominators are not applied retrospectively to those numerators.

Table 31: Historical O0 screen ratios (two datasets and two seeds), retained with their original denominator. These are not the fresh matched-calibration results in Table 6; T0 is the exact control.

Pair	Setting	n	D_{loss}	D_{logit}	D_{pred}
O0	AdamW eps 10^{-8}	4	4.9189	0.7599	1.6390
O0	AdamW eps 10^{-2}	4	0.1345	0.0008	0.0729
T0	AdamW eps 10^{-8}	4	0.0000	0.0000	0.0000

G THRESHOLD SENSITIVITY

Table 32 recomputes the frozen raw divergence rule under modest threshold perturbations. For multiplier α, the parameter and hidden-disagreement thresholds become 0.05α and the validation-loss threshold becomes 0.001α. The main separation between divergent pairs and stable controls is unchanged: stress-protocol AdamW A0/A1/N0/O0 remains divergent, while T0 and SGD controls remain stable. The TinyStories results under 10^{-2} remain a reported boundary case and are interpreted separately from the main stress results.

Table 32: Threshold sensitivity for the 5000-step long-confirmation matrix. Counts are recomputed from raw metrics rather than copied from stored labels.

Mult.	Param/hid	Loss	Default div	T0 stable	SGD stable	Wiki eps	Tiny eps	Core
0.5	0.025	0.0005	40/40	30/30	50/50	20/20	3/20	yes
1	0.05	0.001	40/40	30/30	50/50	20/20	8/20	yes
2	0.1	0.002	40/40	30/30	50/50	20/20	13/20	yes

H Reproducibility and Artifact Release

The supplementary artifact supports auditing the main tables from frozen CSV and report files. The release notes in `README_ARTIFACT_RELEASE.md` specify the package layout, report-generation commands, unit-test entry point, and full training rerun scripts. The A100 return contains 3,078 manifest-verified files: frozen Phase23/24 code snapshots, preregistrations, all run summaries, the complete attempt ledger, and corrected pair-seed prediction outputs. Later additions include the frozen early decision rule, accepted-A0 preregistration, 50,000-step path-uniform extension, five-seed 4090 355M replication, V100 five-seed and horizon reports, manifest-verified Phase30+Phase33 6000D 1.003B A0/T0 aggregate, and independently re-aggregated 6000D 354.9M lower-learning-rate 20,000-step study. All hardware groups remain separate during aggregation. The public package excludes private hostnames, passwords, and account-specific launch state; repository URLs should be added only when venue policy permits de-anonymization.

I Default Screening Matrix Excerpt

The broad screen records supported and unavailable protocol cells, complementing the five-seed confirmations in the main text. The excerpt below contains the original $n = 2$ O0/T0/SGD results used to generate Figure 5.

Table 33: Default-screen O0/T0 excerpt. Values are means over verified real-data rows; divergence follows the pre-specified raw label.

Dataset	Pair	Optimizer	n	Divergent	Mean param rel	Mean val gap
TinyStories	O0	AdamW default	2	2	0.3822	0.0089
TinyStories	O0	SGD	2	0	0.0000	0.0000
TinyStories	T0	AdamW default	2	0	0.0000	0.0000
WikiText2	O0	AdamW default	2	2	0.3870	0.0190
WikiText2	O0	SGD	2	0	0.0000	0.0000
WikiText2	T0	AdamW default	2	0	0.0000	0.0000

J Additional Implementation and Protocol Details

J.1 Implementation Pairs

The main divergence matrix contains five implementation pairs. **A0** compares the public Math and Flash backends of `scaled_dot_product_attention`, forced with `SDPBackend.MATH` and `SDPBackend.FLASH_ATTENTION`. **A1** compares an explicit $QK^\top$/causal-mask/softmax/PV composition with Flash SDPA. **N0** compares `torch.nn.LayerNorm` with an explicit composition of mean, biased variance, reciprocal square root, and affine transformation, using matched epsilon. **O0** compares `torch.optim.AdamW(fused=False, foreach=False)` with `AdamW(fused=True)` under identical hyperparameters. **T0** compares matrix multiplication with `einsum` under the same mask and softmax, providing an exact stable control.

The boundary calibration adds **L8** and **L4**. Both branches share the native LayerNorm forward computation and identical input gradients; the affine gradients at all five LayerNorm sites use fixed eight-chunk versus four-chunk FP32 reduction trees. Their global local-gradient gaps are near 6×10^{-9}, while their affine gradients are bitwise non-exact. The prospective calibration also freezes **T2** (eager versus `torch.compile`/TorchInductor), **A4** (query pre-scaling versus score post-scaling), **N1** (native LayerNorm versus explicit `var_mean(correction=0)`), and **N2** (otherwise identical explicit LayerNorm graphs using `rsqrt` versus reciprocal `sqrt`).

Together, these pairs cover production API replacements (A0/O0), auditable compositions (A1/N0/A4/N1/N2/L8/L4), and a compiler replacement (T2), with T0 calibrating exact alignment. The local gaps measure finite-precision agreement for each pair.

J.2 Full Experimental Protocol

Models and data. We evaluate controlled Transformer language models (Vaswani et al., 2017) on TinyStories (Eldan & Li, 2023) and WikiText2 (Merity et al., 2017). The primary confirmation uses five seeds and 5,000 training steps, including A0 under default AdamW, 10^{-2}, and SGD. A preregistered boundary calibration evaluates L8/L4 under BF16 AdamW at learning rates 3×10^{-3}, 3×10^{-4}, and 10^{-4} with disjoint seed sets. A controlled scaling study repeats O0/T0/N0 at dimension 512 with 8 heads and 6 layers (Appendix Table 20).

For larger models, a memory-bounded sequential runner evaluates 354,923,520-parameter A0 on an RTX 4090 and 1,003,484,160-parameter A0 on an independent RTX 6000D, each with five 5,000-step stress seeds. The latter includes five 1,000-step T0 controls. A separate frozen RTX 6000D study evaluates 354,923,520-parameter A0 at learning rate 10^{-4} for 20,000 steps using fresh seeds 191, 193, and 197. Checkpoints are 1, 50, 100, 200, 500, 1k, 2k, 5k, 10k, 15k, and 20k, with separate terminal and path-uniform decisions. Independent V100 FP16 tests are reported in Appendix Table 28.

The prospective decision study freezes a step-50 rule on existing TinyStories runs. After rejection under the stress protocol, a tuning-only two-seed grid selects learning rate 10^{-4}. A preregistered five-seed confirmation then evaluates the base-model 5,000-step terminal contract. Later preregistered extensions evaluate all declared checkpoints through 50,000 base-model steps and 20,000 355M steps, retaining the original thresholds.

Table 34: Frozen main-study training protocols. The independent V100 protocol is reported in Appendix Table 28.

Setting	Base confirmation	Scale transfer	Large stress	355M low-LR extension
Parameters; width; layers	0.478M; 128; 2	19.23M; 512; 6	354.92M/1.003B; 1280/1792; 18/26	354.92M; 1280; 18
Heads; sequence; effective batch	4; 128; 16	8; 128; 16	20/28; 128; 16	20; 128; 16
Learning rate; schedule	0.003/10^{-4}; constant	0.003; constant	3×10^{-4}; constant	10^{-4}; constant
Steps; seeds	5,000; 5	3,000; 2	5,000; 5 (1B T0: 1,000; 5)	20,000; 3
Adam betas; weight decay	$(0.9, 0.999)$; 0.01	$(0.9, 0.999)$; 0.01	$(0.9, 0.999)$; 0.01	$(0.9, 0.999)$; 0.01
Precision; dropout	BF16; 0	BF16; 0	BF16; 0	BF16; 0

All runs use a byte-level vocabulary of size 256, epsilon 10^{-8} unless specified otherwise, and no gradient clipping. The A100/4090/6000D protocols use BF16 CUDA autocast without GradScaler; the V100 replication uses FP16 autocast. We seed Python, Torch, and CUDA and fix the data order for each seed. Deterministic-algorithm enforcement is disabled because the production kernels are part of the implementation comparison. The base, scale-transfer, 355M, and 1B configurations contain 477,696, 19,230,720, 354,923,520, and 1,003,484,160 trainable parameters, respectively. The large-model runner executes one implementation at a time from a common initial state, retaining only the checkpoints required for paired comparison and deleting each temporary checkpoint afterward. In the 355M lower-learning-rate extension, we apply the frozen raw rule at each declared checkpoint. Terminal stability requires a pass at step 20,000, path-uniform stability requires a pass at every checkpoint, and training maturity requires a reference validation-loss reduction of at least 1.0 from step 50 to step 20,000.

Optimizers and interventions. We compare AdamW at 10^{-8} with an unmatched 10^{-2} setting as a sensitivity probe. To investigate the mechanism, we directly synchronize native first- or second-moment state. A separate epsilon ladder ($10^{-4}/10^{-3}/10^{-2}$) matches measured update norms using a disjoint calibration prefix. Additional controls vary SGD, momentum, beta values, weight decay, and optimizer-state dtype, and include a single-step fused/unfused optimizer check. These interventions build on analyses of Adam-style adaptive optimization and decoupled weight decay (Kingma & Ba, 2015; Reddi et al., 2018; Loshchilov & Hutter, 2019); the large-epsilon setting is used for diagnosis rather than as a production recommendation.

Cross-precision controls show that N0 and O0 can diverge under FP32 AdamW as well as BF16. Precision changes the numerical stress, while repeated adaptive optimization is consistently associated with path separation across these pairs.

Metrics. We measure parameter relative ℓ_2 distance, hidden-state CKA, validation-loss gap, logit symmetric KL, prediction agreement, and prediction disagreement along the paired paths. Local checks measure output max/mean differences and gradient relative ℓ_2. Denominator telemetry includes AdamW denominator and second-moment relative ℓ_2, quantiles of $\epsilon/\sqrt{\hat{v}}$, and the fraction of entries whose denominators are dominated by epsilon.

Reproducible aggregation. Tables are generated from frozen CSV rows, run-level `metrics.jsonl` telemetry, or sequential summaries protected by manifests and hashes. The evidence pack contains real-kernel confirmations, native-state interventions, update-matched epsilon calibration traces, corrected prediction manifests grouped by pair and seed, long-confirmation aggregates, seed-normalized scores, threshold recomputations, and hardware-stratified 4090/V100/6000D summaries, including the 50,000-step extension. We combine the frozen three-seed and separately frozen two-seed 1B studies only after verifying that source, data, model, protocol, GPU UUID, Torch, and CUDA match. The 355M 20,000-step results have a separate manifest and were independently re-aggregated after local hash verification.

K MATHEMATICAL STRUCTURE OF THE SUBSTITUTION CONTRACT

This appendix develops the deterministic identities and conditional bounds used in Section 2. The identities apply to a fixed realization of the declared protocol. Stability assumptions are stated explicitly; their constants are not estimated by E1–E3. The empirical decisions continue to use the frozen observables and raw tolerances.

K.1 ACCEPTANCE SETS AND CONTRACT PROFILES

Fix two realized trajectories, their evaluation probes, and a checkpoint grid $\mathcal{S}_T$ with $T \in \mathcal{S}_T$. For strictly positive tolerances, define the normalized checkpoint discrepancy and the two contract scores by

$$q(t;\boldsymbol{\tau}) = \max_{m\in\mathcal{M}} \frac{\delta_m(t)}{\tau_m}, \tag{14}$$

$$R_{\mathrm{term}}(T;\boldsymbol{\tau}) = q(T;\boldsymbol{\tau}), \qquad R_{\mathrm{path}}(\mathcal{S}_T;\boldsymbol{\tau}) = \max_{t\in\mathcal{S}_T} q(t;\boldsymbol{\tau}). \tag{15}$$

For the path-uniform contract, the maximum over observables commutes with the maximum over checkpoints. The set of passing checkpoints is

$$\mathcal{I}_{\boldsymbol{\tau}} = \{t \in \mathcal{S}_T : q(t;\boldsymbol{\tau}) \le 1\}. \tag{16}$$

Conditional on local qualification, terminal acceptance means $T \in \mathcal{I}_{\boldsymbol{\tau}}$, whereas path-uniform acceptance means $\mathcal{I}_{\boldsymbol{\tau}} = \mathcal{S}_T$. This distinction concerns the monitored grid, not unobserved updates between checkpoints.

Proposition 1 (order properties). For fixed trajectories and evaluations, with the same local gate: (i) path-uniform acceptance implies terminal acceptance; (ii) enlarging any tolerance cannot turn an acceptance into a rejection; and (iii) adding checkpoints cannot turn a path-uniform rejection into an acceptance.

Proof. Since $T \in \mathcal{S}_T$, $R_{\mathrm{term}} \le R_{\mathrm{path}}$, proving (i). If $\tau'_m \ge \tau_m > 0$ for every m, then $\delta_m(t)/\tau'_m \le \delta_m(t)/\tau_m$, so both scores weakly decrease, proving (ii). If $\mathcal{S}_T \subseteq \mathcal{S}'_T$, then

$$R_{\mathrm{path}}(\mathcal{S}_T;\boldsymbol{\tau}) \le R_{\mathrm{path}}(\mathcal{S}'_T;\boldsymbol{\tau}), \tag{17}$$

which proves (iii). The shared gate multiplies each indicator by the same Q_{loc} and does not change these implications. □

For a common multiplier $\alpha > 0$, this gives an explicit one-dimensional tolerance profile:

$$R_{\mathcal{C}}(\alpha\boldsymbol{\tau}) = \alpha^{-1} R_{\mathcal{C}}(\boldsymbol{\tau}), \qquad \mathrm{Accept}_\alpha = Q_{\mathrm{loc}}\,\mathbf{1}[\alpha \ge R_{\mathcal{C}}(\boldsymbol{\tau})]. \tag{18}$$

The infimum of accepting multipliers for a locally qualified realized pair is therefore its contract score. This is an algebraic characterization of a fixed trajectory, not a rule for selecting confirmatory tolerances after observing outcomes. Table 27 reports the declared tolerance probes. The report and dense grids obey Proposition 1 because their shared checkpoints use the same recorded evaluations.

Terminal decisions have no corresponding monotonicity in the horizon: $q(T)$ can rise and later fall. In the A0 extension, the same trajectories pass at 5k and 50k while crossing an intermediate loss-gap threshold. Comparisons that also change the model, learning rate, runtime, or probe concern different contracts and are not ordered by Proposition 1.

K.2 PROPAGATION OF IMPLEMENTATION DEFECTS

Use a fixed norm on the augmented state space; if parameter and optimizer-state blocks use different units, their relative weights must be specified and held fixed. Define

$$e_t = z_t^{(c)} - z_t^{(r)}, \qquad \eta_t = F_{c,t}(z_t^{(r)}) - F_{r,t}(z_t^{(r)}). \tag{19}$$

Assume that nonnegative L_t and ε_t satisfy, at each compared step,

$$\|F_{c,t}(z_t^{(c)}) - F_{c,t}(z_t^{(r)})\| \le L_t\|e_t\|, \qquad \|\eta_t\| \le \varepsilon_t. \tag{20}$$

Only this pairwise stability inequality is needed; no differentiability of the finite-precision implementation map is assumed.

Proposition 2 (finite-horizon propagation bound). Under Equation 20, for arbitrary initial separation,

$$\|e_T\| \le \left(\prod_{j=0}^{T-1} L_j\right)\|e_0\| + \sum_{k=0}^{T-1}\left(\prod_{j=k+1}^{T-1} L_j\right)\varepsilon_k, \tag{21}$$

where an empty product is one. Equation 7 follows when $e_0 = 0$.

Proof. Adding and subtracting $F_{c,t}(z_t^{(r)})$ gives Equation 6. The triangle inequality yields $\|e_{t+1}\| \le L_t\|e_t\| + \varepsilon_t$. At $T = 1$ this is Equation 21. Multiplying the bound at T by L_T and adding ε_T appends the factor L_T to every existing product and adds the empty-product term for $k = T$, establishing the result by induction. □

For uniform bounds $L_t \le L$ and $\varepsilon_t \le \varepsilon$, shared initialization gives

$$\|e_T\| \le \varepsilon \sum_{k=0}^{T-1} L^k = \begin{cases} \varepsilon(1-L^T)/(1-L), & 0 \le L < 1, \\ T\varepsilon, & L = 1, \\ \varepsilon(L^T-1)/(L-1), & L > 1. \end{cases} \tag{22}$$

The three regimes separate contraction, linear accumulation, and possible amplification. The upper bound alone does not imply that an empirical trajectory diverges; actual defects can cancel.

Construction (small local defect, large accumulated discrepancy). Consider scalar maps $F_r(x) = ax$ and $F_c(x) = ax + \varepsilon$, with $a \ge 1$, $\varepsilon > 0$, and $x_0^{(r)} = x_0^{(c)} = 0$. At every shared state,

$$|F_c(x) - F_r(x)| = \varepsilon, \qquad |x_T^{(c)} - x_T^{(r)}| = \varepsilon \sum_{k=0}^{T-1} a^k. \tag{23}$$

For $a = 1$, any $T > \tau/\varepsilon$ violates a fixed terminal tolerance τ, despite a uniformly bounded same-state defect at every step. For $a > 1$, the geometric expression is attained exactly. This construction concerns full update maps, giving more information than an output/gradient audit alone. It shows why a small same-state discrepancy, without propagation and horizon information, does not yield a horizon-independent preservation guarantee. It is an analytic illustration, not a model fitted to the reported neural experiments.

From state bounds to observable bounds. Suppose an observable discrepancy additionally satisfies $\delta_m(t) \le K_{m,t}\|e_t\|$ for known nonnegative constants along the compared trajectories. Set $B_0 = \|e_0\|$ and $B_{t+1} = L_t B_t + \varepsilon_t$. Then sufficient conditions are

$$Q_{\mathrm{loc}} = 1, \qquad \begin{cases} K_{m,T} B_T \le \tau_m \quad \forall m, & \text{terminal}, \\ \max_{t \in \mathcal{S}_T} K_{m,t} B_t \le \tau_m \quad \forall m, & \text{path-uniform}. \end{cases} \tag{24}$$

These additional regularity constants are not established for our measured CKA or prediction-disagreement observables. The implementation therefore evaluates those observables directly rather than substituting this sufficient condition for the empirical contract.

K.3 Optimizer Memory and Gradient Feedback

The augmented state makes optimizer memory explicit. For an idealized Adam first-moment update with identical β_1 and matched step counts, let $\Delta m_t = m_t^{(c)} - m_t^{(r)}$ and $\Delta g_t = g_t^{(c)} - g_t^{(r)}$. With zero initial moment separation,

$$\Delta m_t = \beta_1 \Delta m_{t-1} + (1-\beta_1)\Delta g_t = (1-\beta_1)\sum_{k=1}^{t} \beta_1^{t-k} \Delta g_k. \tag{25}$$

In finite-precision native kernels, define the residual $q_t^{(i)}$ by writing the implemented update as $m_t^{(i)} = \beta_1 m_{t-1}^{(i)} + (1-\beta_1) g_t^{(i)} + q_t^{(i)}$. The exact residual-aware identity is then

$$\Delta m_t = \beta_1^t \Delta m_0 + \sum_{k=1}^{t} \beta_1^{t-k}\big[(1-\beta_1)\Delta g_k + \Delta q_k\big]. \tag{26}$$

This distinguishes accumulated gradient disagreement from differences in the moment-update implementation. In O0, the local forward/backward computation is identical at a shared state, while separate optimizer updates can create the parameter separation that generates later Δg_t.

For positive elementwise denominators $d_t^{(i)} = \sqrt{\hat{v}_t^{(i)}} + \epsilon$, the difference in normalized moment terms has the exact componentwise decomposition

$$\frac{\hat{m}_t^{(c)}}{d_t^{(c)}} - \frac{\hat{m}_t^{(r)}}{d_t^{(r)}} = \frac{\Delta \hat{m}_t}{d_t^{(c)}} - \frac{\hat{m}_t^{(r)} \Delta d_t}{d_t^{(c)} d_t^{(r)}}. \tag{27}$$

With common learning rate α_t and weight decay λ, an idealized decoupled-weight-decay update therefore gives

$$\Delta\theta_{t+1} = (1 - \alpha_t \lambda)\Delta\theta_t - \alpha_t \left(\frac{\Delta \hat{m}_t}{d_t^{(c)}} - \frac{\hat{m}_t^{(r)} \Delta d_t}{d_t^{(c)} d_t^{(r)}} \right). \tag{28}$$

Native finite-precision parameter updates add their own residual difference. The first-order expression in Section 6 is the adaptive-step component obtained by expanding around a common denominator. Equations 26–27 explain the terms targeted by the native-state interventions; the measured intervention results, rather than these identities alone, support the first-moment-dominant interpretation within O0.

K.4 Screening Errors and the Calibration Estimand

For the E2 cohort $\mathcal{U}$, write $y_u \in \{0, 1\}$ for final raw-contract rejection and $\widehat{y}_u \in \{0, 1\}$ for predicted rejection. The confusion counts are

$$\begin{pmatrix} \mathrm{TP} & \mathrm{FN} \\ \mathrm{FP} & \mathrm{TN} \end{pmatrix} = \sum_{u \in \mathcal{U}} \begin{pmatrix} y_u \widehat{y}_u & y_u(1 - \widehat{y}_u) \\ (1 - y_u)\widehat{y}_u & (1 - y_u)(1 - \widehat{y}_u) \end{pmatrix}. \tag{29}$$

Here $|\mathcal{U}| = 35$, $\sum_u y_u = 20$, and $\sum_u (1 - y_u) = 15$. The paired rule gives $(\mathrm{TP}, \mathrm{TN}, \mathrm{FP}, \mathrm{FN}) = (20, 15, 0, 0)$; the cumulative step-200 local policy gives $(17, 10, 5, 3)$. These are the existing results of Table 1. Five seed IDs are shared across conditions, so the matrix is a cohort summary rather than 35 independent-seed trials.

The matched diagnostic denominator in Equation 10 averages all ten unordered pairs of five reference trajectories. If $h_m(a, b) = g_m^{r_a, r_b; \mathrm{probe}}$, then its variance, when reference runs are treated as random and second moments exist, contains covariance terms:

$$\mathrm{Var}(b_m) = \binom{K}{2}^{-2} \sum_{a<b} \sum_{u<v} \mathrm{Cov}(h_m(a, b), h_m(u, v)). \tag{30}$$

Pairs sharing a reference run need not be independent; ten pairwise gaps therefore do not constitute ten independent calibration runs. The ratio D_{TPE}^m expresses the paired gap relative to a mean initialization-variation scale, with a 10^{-12} denominator safeguard. It is not a tail probability or a confidence bound. Both numerator and denominator use the common probe, while the raw binary outcome remains tied to the separately retained legacy evaluation. No historical numerator is renormalized using the fresh denominators.