RESEARCH RECORD · 11 SEPTEMBER 2026

Adaptive Consensus

From Diagnostics to Training Dynamics

This supplement records the recipes, mathematical definitions, historical implementation, and limitations behind the 13-slide presentation. All empirical claims concern saved runs and analyses in this repository. The proposed fresh-update method has frozen-state checks but **was untested in training at the review stage** in this investigation. The scope ends at the requested `consensus-scaling-review-2026-09-11` record. A later `fresh-update-trial-2026-09-11` campaign is present in the workspace but is not included here; "missing control" and "untested" refer to the evidence available in the requested review.

Main finding. Moment weighting raises the directional-curvature statistic on ordinary adaptive-consensus checkpoints. Repeatedly applying the weighting to persistent parameter disagreement did not improve the available training results and produced severe coordinate concentration. A favorable checkpoint diagnostic is therefore insufficient evidence for the resulting training dynamics.

1 Experimental recipe

1.1 Model, data, and training budget

Each local model has **20,403,520 unique parameters**: an eight-block, decoder-only Llama-style transformer with width 320, five attention heads, head dimension 64, and feed-forward width 896. Vocabulary size is 32,000 and context length is 1,024. Input embeddings are tied to the output projection and counted once. The model uses rotary positions, RMS normalization, and a gated feed-forward block. [S1–S6; S8]

Setting	Value
Initialization / normalization	Normal initialization std. 0.02; RoPE base 10,000; RMS epsilon 10^{-5}
Dataset	C4, `allenai/c4`; tokenizer `TinyLlama/TinyLlama-1.1B-intermediate-step-1431k-3T`
Budget	20 tokens / parameter; approximately 408M global training tokens; 40 reporting epochs
Epoch convention	0.5 tokens / parameter per reporting epoch; this is a token-budget unit, not a full pass over C4
AdamW	$\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$; weight decay 0.1 on tensors with at least two dimensions; zero decay on normalization vectors
Gradient clipping	Per-local-model global L2 norm clipped at 1.0 before mixing
Schedule	5% linear warmup; cosine decay to 10% of peak LR
Runtime	Seed 42; nondeterministic kernels; one NVIDIA GH200; CUDA; BF16 AMP, compiled loss, SDPA attention, fused AdamW; eight CPU threads

Global batches contain 32,768 tokens. Synchronous training updates one model from 32 sequences at peak LR 0.0022. Packed training evaluates four independent local models in a leading worker dimension on one GPU; each receives eight sequences, or 8,192 tokens, and has separate Adam moments at peak LR 0.0048. Local losses are summed for backpropagation so each worker receives its own local-mean gradient without an extra factor of four. Every fourth sample in the shared buffered stream is assigned to the same worker.

The experiment therefore represents decentralized optimization within one device; it does not measure inter-device communication cost or scaling. [S7, S8]

The synchronous run completed 408,071,168 tokens and 12,480 steps. All packed comparison runs completed 408,068,096 global tokens, 102,017,024 tokens per local model, and 12,473 steps. Slightly different rounding and shortened steps at reporting-epoch boundaries account for these differences. Packed reporting epochs normally contain 10,207,232 tokens. [S1–S6]

1.2 Exact remaining recipe fields and evaluation

Field group	Resolved settings
Data revision	`1588ec454efa1a09f29cd18ddd04fe05fc8653a2`
Tokenizer revision	`59f6f375b26bde864a6ca194a9a3044570490064`
Data pipeline	Cache `data/c4`; shuffle seed 42; shuffle buffer 10,000; shard tokens 16,777,216; tokenization batch 256; buffer 64 MiB; prefetch enabled; preparation limits null
Logging / checkpoints	Log every 20 steps; checkpoint interval 500; policy `all` for the four principal recipes; epoch training state saved. Review variants may differ in checkpoint retention and output path
Validation	Subset: 1,024 blocks, batch 128, seed 12,345. Final full validation: 197,411,295 scored tokens
Runtime details	Device `cuda:0`; compile mode `default`; SDPA backend `auto`; deterministic false; output directory is the run identifier
Synchronous budget fields	`max_tokens=null`; `epoch_tokens=null`; derived from parameter count and tokens-per-parameter settings
Packed budget fields	`max_tokens=408068096`; `epoch_tokens=10207232`

Packed evaluation copies the **arithmetic mean of local parameters** into a single evaluation model; it does not average worker logits or losses. Subset-validation trajectories and final full-validation losses use different data coverage and must remain separately labeled. All main-deck diagnostic plots use the unseen analysis case, defined below; that case is also distinct from validation. Exact principal YAML recipes are preserved verbatim in `supplement-data.json`; archived per-run recipes and hashes are listed in the package provenance. [S1–S6]

Let c_t be consumed tokens after the step, C the full token budget, $w = 0.05C$, and $\eta_{\max}$ the configured peak LR. The executed schedule is

$$\eta_t = \begin{cases} \eta_{\max}\frac{c_t}{w} & c_t < w \\ \eta_{\max}\left[0.1 + 0.9\frac{1+\cos\left(\pi\frac{c_t-w}{C-w}\right)}{2}\right] & c_t \geq w \end{cases}. \tag{1}$$

The implementation bounds cosine progress to $[0, 1]$. The active adaptive-consensus LR maximum is computed separately over the exact planned steps, including shortened epoch steps. [S8]

2 The executed decentralized algorithms

2.1 Topology and adaptive mixing

Write $\theta_{i,t}$ for worker i before the step, W_t for its row-stochastic mixing matrix, and $b_{i,t} = \sum_j W_{ij,t}\theta_{j,t}$ for the ordinary mixing anchor. For four-worker **complete topology**, $W_{ij,t} = \frac{1}{4}$ and $b_{i,t} = \overline{\theta}_t$. For **one-peer exponential topology**, worker i averages itself with worker $i - 2^{t \bmod 2}$ modulo four: offsets alternate between one and two. Each update averages two parameter vectors; it is not full averaging on each step. [S8]

The initial plain-versus-adaptive comparison uses one-peer exponential topology. The moment-scaled review runs use complete topology. Full mixing means $\gamma_t = 1$ on the specified topology; the complete-

topology full-mixing baseline averages all four parameters every step. Even with complete mixing, local Adam moments and local updates remain different, so this baseline is not synchronous gradient averaging.

For ordinary adaptive consensus (AC),

$$b_{i,t} = \sum_j W_{ij,t}\theta_{j,t}, \quad y_{i,t} = \gamma_t b_{i,t} + (1-\gamma_t)\theta_{i,t}, \tag{2}$$

$$\gamma_t = \begin{cases} 1 & t < t_0 \\ \left(\frac{\eta_t}{\eta_{\max,\text{active}}}\right)^p & t \geq t_0 \end{cases}, \quad t_0 = \lceil 0.25T \rceil, \quad p = 2. \tag{3}$$

Here t is zero-based, $T = 12,473$, and $t_0 = 3,119$. The denominator is $\max_{s \geq t_0} \eta_s$, the largest LR **inside the active interval**, not the configured peak LR. At the first active step the ratio is one, so mixing initially remains unchanged; weakening begins subsequently. At epoch 10, the saved last-step γ is still one; review replays reuse that saved value. The final saved replay value is approximately 0.01221 for $p = 2$. The $\alpha = 0.025$ variant instead uses $p = 1$, with final value approximately 0.11050. [S8; S6]

2.2 Optimizer equations and ordering

Gradients are evaluated at each worker's **pre-mixing parameters**. Let $g_{i,t}$ be the local-mean gradient and $\tilde{g}_{i,t}$ its clipped version,

$$\tilde{g}_{i,t} = g_{i,t} \min\left(1, \frac{1}{\|g_{i,t}\|_2 + 10^{-6}}\right). \tag{4}$$

The additive 10^{-6} reflects the clipping coefficient convention; ordinary mathematical global-norm clipping gives the same limiting interpretation. The logged local gradient norms are measured **before** this clipping.

The code then changes parameters to $y_{i,t}$ by mixing and calls the independent AdamW optimizers, which update local moments using those already computed gradients:

$$m_{i,t+1} = \beta_1 m_{i,t} + (1-\beta_1)\tilde{g}_{i,t}, \tag{5}$$

$$v_{i,t+1} = \beta_2 v_{i,t} + (1-\beta_2)\tilde{g}_{i,t}^2, \tag{6}$$

$$\hat{m}_{i,t+1} = \frac{m_{i,t+1}}{1-\beta_1^{t+1}}, \quad \hat{v}_{i,t+1} = \frac{v_{i,t+1}}{1-\beta_2^{t+1}}, \tag{7}$$

$$u_{i,t} = -\eta_t \frac{\hat{m}_{i,t+1}}{\sqrt{\hat{v}_{i,t+1}} + \varepsilon}, \quad \theta_{i,t+1} = (1 - \eta_t\lambda) \odot y_{i,t} + u_{i,t}. \tag{8}$$

All squares, divisions, and decay coefficients are coordinatewise. The vector λ is 0.1 for matrix-like parameters and zero for one-dimensional parameters. The key order is: **local forward/backward → local clipping → parameter mixing → local AdamW moments and displacement**. Gradient clipping does not clip the separate consensus transformation. [S7, S8]

2.3 Tested moment-scaled consensus error

Define **consensus error** as $\delta_{i,t} = \theta_{i,t} - \overline{\theta}_t$, the difference between a local model and the global average model. In the tested complete topology, $b_{i,t} = \overline{\theta}_t$. The transformation starts from the consensus error that ordinary AC would retain:

$$e_{i,t} = (1-\gamma_t)\big(\theta_{i,t} - b_{i,t}\big), \quad s_{i,t} = e_{i,t} \odot v_{i,t}^{\alpha}, \tag{9}$$

$$y_{i,t} = b_{i,t} + \left(\frac{\|e_{i,t}\|_2}{\|s_{i,t}\|_2}\right) s_{i,t}. \tag{10}$$

It uses the **previous-step raw local second moment** $v_{i,t}$, before the current gradient updates it. The L2 normalization spans all 20,403,520 unique parameters of each worker; it is not per tensor or per layer. Tied embedding/output weights appear once, and padding is zero and excluded by the active-coordinate mask. Principal settings are complete topology, $\alpha = 0.5$, $p = 2$, and activation fraction 0.25. [S6; S8]

For $\alpha = 0$, the implementation follows ordinary AC. For $\gamma = 1$, it performs ordinary full mixing and skips consensus error scaling. If the consensus error or its scaled direction is zero, it retains the original consensus error. For negative α, moments are floored at 10^{-16}; positive powers use raw nonnegative moments. Invalid or negative moments are rejected. To avoid overflow, the executed helper forms the direction with centered log weights and reorients the retained error in FP64, then copies the result to FP32 parameter storage. The reweighted vector preserves $\|e_{i,t}\|_2$ about the pre-mixing global average. Worker-dependent reweighting can shift the average, so the actual post-transformation consensus error is $y_{i,t} - \left(\frac{1}{4}\right)\sum_j y_{j,t}$; its norm need not equal $\|e_{i,t}\|_2$. [S8]

> **Historical source matters.** The named branch `feature/adaptive-consensus-moment-scaling` points to commit `26a3bddb734b4d72b7a7ec287280be18b8fe804a`; that committed tree contains ordinary adaptive consensus but lacks the committed scaling implementation. The equations above come from the preserved executed source under the scaled run's `analysis/alignment/source/`, together with its resolved recipe. A mutable current checkout is not used to establish what these runs executed.

3 Diagnostic definitions and numerical limits

3.1 Data cases and gradient noise

Diagnostics use nine saved epochs: 1, 5, 10, 15, 20, 25, 30, 35, and 40. The **seen** case replays the first reporting-epoch-sized training stream with the training loader seed. The **unseen** case draws a separate epoch-sized slice immediately beyond the entire consumed training range, with a derived analysis shuffle seed. Both come from the **training cache**. Unseen does not mean the C4 validation split or a separate downstream task. [S7]

For packed runs each case contains 9,968 blocks, or 10,207,232 tokens. The unseen range begins at training-cache block 398,504. Synchronous cases contain 9,963 blocks, or 10,202,112 tokens, with unseen start block 398,507. At each checkpoint, derivatives are evaluated at the single saved evaluation model (the averaged model for packed runs), not at four different local points.

Let L_D be mean token loss over this fixed diagnostic epoch, $\bar{g} = \nabla L_D\left(\bar{\theta}\right)$, and $H = \nabla^2 L_D\left(\bar{\theta}\right)$. For 32 sampled local batches B_k, selected without replacement using a fixed analysis seed,

$$n_k = \nabla L_{B_k}\left(\bar{\theta}\right) - \bar{g}, \quad N = \left(\frac{1}{32}\right)\sum_{k=1}^{32}\|n_k\|_2, \quad G = \|\bar{g}\|_2. \tag{11}$$

The plotted noise-to-mean ratio is $\frac{N}{G}$, using an average of norms rather than the square root of mean squared norm. Packed noise batches preserve the 8,192-token local batch and interleaved-worker sampling; synchronous noise batches use 32,768 tokens. Short final batches remain eligible; the saved samples here all use the stated full batch sizes. The mean gradient is computed over the whole diagnostic epoch, not estimated from the 32 noise samples. [S7]

Unseen checkpoint	Mean-gradient norm	Mean noise norm	Noise / mean
Synchronous, epoch 1	0.304839	0.390946	1.2825
Synchronous, epoch 40	0.113477	0.296089	2.6092
Packed plain, epoch 1	0.225797	0.627213	2.7778
Packed plain, epoch 40	0.109880	0.704206	6.4089

The endpoint ratios support increasing **relative** noise dominance within these runs; the trajectory is not strictly monotone, and absolute noise need not increase. Cross-run magnitudes are affected by distinct local batch sizes, LRs, and trajectories. These are raw loss-gradient diagnostics; they do not directly measure noise after Adam preconditioning.

3.2 Directional curvature ("normalized alignment")

For a collection of directions $D = \{d_i\}$, define

$$A(D; H) = \frac{\sum_i d_i^T H d_i}{\sum_i \|d_i\|_2^2}. \tag{12}$$

For noise, use $d_i = n_i$. For consensus, use $\delta_i = \theta_i - \bar{\theta}$, a displacement from the **full worker mean** even when training uses exponential topology. The random reference uses 32 independent Rademacher vectors in dimension $P = 20,403,520$ and divides the sum of their quadratic forms by $32P$, giving a stochastic reference for $\frac{\mathrm{tr}(H)}{P}$. [S7]

This statistic is a pooled Rayleigh quotient, with curvature units; it is neither cosine alignment nor a fraction restricted to $[0, 1]$. "Better alignment" in the slides means larger directional curvature. The original files also report unnormalized $\mathrm{mean}_i(d_i^T H d_i)$ and average direction norm; a large normalized value can coexist with a tiny raw quadratic form after moment multiplication.

The scaled diagnostic forms seven direction families: δ, δ/m, $\delta \odot m$, δ/v, $\delta \odot v$, $\delta/\sqrt{v}$, and $\delta \odot \sqrt{v}$. Raw Adam moments are used without bias correction. Division by m preserves sign and floors absolute magnitude at 10^{-12}; division by v floors it at 10^{-16}; the square-root denominator floor is 10^{-8}. Multiplications apply raw moments without floors. [S4]

Ordinary AC, epoch 20	Seen A	Unseen A
δ	0.00111285	0.00111710
$\delta \odot \sqrt{v}$	0.0455245	0.0454412
$\delta \odot v$	2.44631	2.43292

The unseen increases are 40.68× and 2,177.89× relative to the unscaled consensus error. They are checkpoint-direction comparisons, not comparisons of trained algorithms. On the **scaled training trajectory** at epoch 20, unseen values instead are 0.00194702 for δ and 3.28126×10^{-6} for $\delta \odot \sqrt{v}$. The diagnostic signal from the baseline does not persist under the modified dynamics. [S3, S6]

3.3 Precision and similarity measurements

Hessian-vector products use FP32 parameters, BF16 forward AMP, TF32, reference attention, and compiled Pearlmutter differentiation with batches of 64 sequences. The scaled diagnostic forms and norms directions in FP64, normalizes and casts them to FP32 for HVPs, and rescales the resulting quadratic forms by the original squared norms. Precision comparisons are informational: the saved seven-family campaign has **no precision acceptance threshold**. Its historical reports include an 8.39% discrepancy for the baseline $\delta \odot m$ check. Small values and sign changes require further precision checks; the slides do not treat them as definitive evidence of positive or negative curvature. [S7, S4]

The full-state similarity analysis computes cosine and symmetric relative L2 distance over concatenated unique parameters:

$$C(a, b) = \frac{a^T b}{\|a\|_2 \|b\|_2}, \quad D(a, b) = \frac{2\|a - b\|_2}{\|a\|_2 + \|b\|_2}. \tag{13}$$

It reports means and population standard deviations over six unordered worker pairs. Two zero vectors have relative distance zero; undefined zero-norm cosines are omitted. "Adam direction" means bias-corrected $\frac{m}{\sqrt{v}+\varepsilon}$, with bias correction applied inside each moment, excluding LR and weight decay. Original similarity

reductions use FP32 with AMP and TF32 disabled; the later review recomputes checkpoint diagnostics in FP64. [S5; S6]

Ordinary AC, epoch 40	Mean cosine	Mean relative L2
Parameters	0.997046	0.0767827
First moment	0.010415	1.40706
Second moment	0.998145	0.0660809
Adam direction	0.003653	1.41216

The high second-moment cosine supports similar directions among the four workers; it does not establish identical magnitudes, identical coordinates, or similar local optimizer steps. Pair spreads describe this group of workers and are not independent-seed confidence intervals.

4 Training results and comparison limits

The original exponential-topology runs finish at full-validation loss 3.595877 for plain decentralized mixing and 3.606162 for AC: a difference of +0.010285 for AC. This is one seed with nondeterministic execution, so it is evidence of **no observed benefit in this comparison**, not a significance estimate. [S2, S3]

The following table reproduces all twelve rows in the review CSV. "Excluded" means embeddings always receive ordinary mixing while the remaining parameters receive AC. Every row has four local models, seed 42, LR 0.0048, $\beta_2 = 0.999$, and 408,068,096 global training tokens. Perplexity is $\exp(\text{loss})$. Maxima use logged steps after 3,119, the AC activation boundary, for every run: training-loss windows and pre-clipping gradient norms; they are not full-validation losses. [S6]

Run	p	Full loss	PPL	Max loss	Max $\|g\|$
Plain / exponential	—	3.595877	36.448	4.066	1.171
AC / exponential	2	3.606162	36.824	4.072	1.348
AC excl. / exponential	2	3.607998	36.892	4.074	1.173
AC excl. / complete	2	3.588548	36.182	4.055	1.204
Scaled α=0.5	2	3.600173	36.605	34.179	1664.4
Scaled α=0.1 (t2)	2	4.026407	56.059	4.846	391.98
Scaled α=0.05 (t3)	2	3.599727	36.588	4.176	39.108
Scaled α=−0.25 (t4)	2	3.583167	35.987	4.058	1.436
Scaled α=−0.5 (t5)	2	3.582657	35.969	4.061	1.259
Scaled α=−1 (t6)	2	3.589681	36.223	4.058	1.296
Scaled α=0.025 (t7)	1	3.594879	36.411	4.054	1.475
Full mixing / complete	—	3.579905	35.87	4.064	1.288

All scaled rows use **complete topology** and include embeddings. The complete full-mixing baseline is best at 3.579905. The best negative-power result, $\alpha = -0.5$, is 3.582657 (+0.002752); $\alpha = -0.25$ gives 3.583167. Principal positive scaling, $\alpha = 0.5$, gives 3.600173 (+0.020267); $\alpha = 0.1$ gives 4.026407 (+0.446501). The $\alpha = 0.025$ run also changes p to one, so it does not isolate the effect of shrinking α.

Missing control. Within the requested review there is no saved complete-topology, all-parameter ordinary AC run with $\alpha = 0$, activation fraction 0.25, and $p = 2$. The exponential AC run changes topology; complete AC with excluded embeddings changes the parameter subset. Neither is a matched control for consensus error scaling. Full mixing remains the strongest existing complete-topology reference.

5 Concentration and worker-mean drift

5.1 Observed trajectory

The review reads fourteen saved checkpoints and recomputes consensus errors and optimizer statistics in FP64 over all unique parameters. Define consensus error concentration as

$$K = \max_i \max_j \frac{\delta_{i,j}^2}{\sum_l \delta_{i,l}^2}. \tag{14}$$

This is the worst worker's largest-coordinate share of its consensus error energy, not a share of model parameter energy. In the principal $\alpha = 0.5$ run it rises as follows. [S6]

Epoch	Mean v cosine	Largest energy share	Mean $\|\delta\|$
10	0.998051	0.00109%	3.3507
15	0.982122	7.61844%	3.1764
20	0.945656	31.05558%	2.9787
24	0.723815	54.10758%	2.2084
26	0.963981	74.47251%	2.683
40	0.911078	95.66187%	2.1662

At epoch 24, the second-moment cosine of block 7 is 0.683319, while blocks 0–5 remain above 0.99. Subset validation rises from 3.82326 at epoch 23 to 4.03620 at epoch 24 and 4.23185 at epoch 26. The largest logged local gradient norm after activation is 1,664.4 before clipping, and a logged training-loss window reaches 34.18. The dominant consensus error coordinate lies in `blocks.7.ffn.down_proj.weight` at epoch 20 and `blocks.7.ffn.up_proj.weight` at epoch 40. These observations support a concentration mechanism but do not isolate every causal link.

5.2 Why norm preservation does not control direction

For two symmetric workers, complete mixing, fixed positive diagonal moment weights, and no optimizer forcing, write the repeated consensus error map as $e^{k+1} = c_k e^k \odot v^\alpha$, where c_k combines scalar mixing contraction and normalization. For any nonzero coordinate pair,

$$\frac{e_j^k}{e_l^k} = \frac{e_j^0}{e_l^0}\left(\frac{v_j}{v_l}\right)^{k\alpha}. \tag{15}$$

Both common scalars cancel. With $\frac{v_j}{v_l} = 10$ and $\alpha = 0.1$, 100 applications amplify the ratio by 10^{10}. The full L2 norm can remain prescribed while the direction approaches a single coordinate. This is an exact illustration under stated assumptions, not a convergence theorem for the nonlinear, evolving AdamW system. Positive powers select larger-moment coordinates; negative powers select smaller-moment coordinates. Either can concentrate error.

5.3 Frozen-state mean-drift counterfactuals

Local weights and per-worker normalization need not preserve the mean: even if $\sum_i e_i = 0$, generally $\sum_i c_i(e_i \odot v_i^\alpha) \neq 0$. In a frozen complete-mixing replay of epoch 40 with $\alpha = 0.5$, mean drift is approximately 0.1155 in parameter-space L2, or 46% of the reconstructed mean Adam displacement norm. The reconstruction uses the saved LR and moments, excluding weight decay; it is not the next logged training update. Negative-power runs also exhibit drift, so drift alone does not explain their better losses. [S6]

Sharing v is insufficient if the normalizer differs across workers. Take $e_1 = (1, 0)$, $e_2 = (0, 1)$, $e_3 = (-1, -1)$ and shared gains $d = (2, 1)$. Workerwise norm preservation gives

$$r_1 = (1, 0), \quad r_2 = (0, 1), \quad r_3 = \sqrt{\frac{2}{5}}(-2, -1), \tag{16}$$

whose average is $(-0.0883037, 0.122515)$. A single pooled normalization factor would preserve zero mean. Shared moments also fail to prevent concentration: at epoch 20, local $\sqrt{v}$ places up to 99.331% of the transformed error energy in one coordinate; shared $\sqrt{\overline{v}}$ still places 99.157% there. This replay uses the full worker average as the anchor; on the exponential baseline it is a hypothetical complete-topology transform.

6 Proposed fresh-update shaping experiment

Status at review stage: untested in training. The following is a specified next experiment, supported by algebra and frozen-state measurements only. No validation-loss result in the requested review evaluates it.

Keep ordinary AC responsible for contracting existing parameter disagreement. Let $u_{i,t}$ be the fresh local Adam adaptive displacement from the current clipped gradient, excluding decoupled weight decay, and let $z_{i,t}$ be the ordinary resulting AdamW parameter. Define

$$q_{i,t} = u_{i,t} - \overline{u}_t, \quad \overline{v}_t = \left(\frac{1}{4}\right)\sum_i v_{i,t}. \tag{17}$$

The shared moments are the preceding-step raw local moments, used only for shaping. Local AdamW state remains independent and unchanged. For each unique parameter tensor $\mathcal{T}$, choose shared bounded gains

$$d_j = \mathrm{clip}\left(\left(\frac{\overline{v}_{t,j} + 10^{-16}}{\mathrm{mean}_{l\in\mathcal{T}}\overline{v}_{t,l} + 10^{-16}}\right)^{\frac{1}{2}}, 0.5, 2\right). \tag{18}$$

Let Q stack every q_i across all workers and unique parameters. Set

$$S_i = q_i \odot d, \quad R_i = \left(\frac{\|Q\|_F}{\|S\|_F}\right) S_i, \tag{19}$$

$$z_i' = z_i + \rho_t(R_i - q_i), \quad \rho_t = 0.1(1-\gamma_t). \tag{20}$$

Skip correction when $Q = 0$. Shared gains in $[0.5, 2]$ ensure a nonzero S whenever Q is nonzero. These bounds, exponent 0.5, tensorwise gain reference, and maximum shaping fraction 0.1 are initial defaults, not tuned optima.

6.1 Algebraic checks and interpretation

Because the gain vector and normalization are shared, $\sum_i q_i = 0$ implies $\sum_i R_i = 0$. The correction therefore preserves the worker-average AdamW update exactly in real arithmetic. Its fresh disagreement is $Q' = (1-\rho)Q + \rho R$, where $\|R\|_F = \|Q\|_F$. By the triangle inequality,

$$\|Q'\|_F \le (1-\rho)\|Q\|_F + \rho\|R\|_F = \|Q\|_F, \quad 0 \le \rho \le 1. \tag{21}$$

This pooled bound applies to the **fresh optimizer displacement**, not the accumulated consensus error after the step. The correction vanishes for $\gamma = 1$ and shrinks with LR through u. Shared bounded gains and a common normalizer avoid the identified mean-drift mechanism. The method does not repeatedly multiply accumulated parameter error by v^{α}.

Frozen-state comparisons reconstruct update disagreement from saved moments at the same epoch; they do not simulate a new batch or the proposed one-step moment lag. Entries below are worst-worker coordinate-energy shares **after hypothetical shaping and before the 0.1 interpolation**. The accumulated-error column uses local weights; fresh-update columns use shared weights. At epoch 10, actual correction is zero because $\gamma = 1$. [S6]

Epoch	Accumulated error	Fresh / unbounded	Fresh / bounded
10	0.4303%	0.4276%	0.00446%
15	99.6701%	3.6059%	0.00456%
20	99.331%	0.1612%	0.00385%
24	98.8564%	28.7946%	0.03692%
40	99.9975%	0.0838%	0.00335%

6.2 Acceptance criteria for a future training campaign

First run the missing complete-topology, all-parameter $\alpha = 0$ AC control with activation fraction 0.25 and $p = 2$. Compare the proposed method with that control, complete full mixing, and the existing negative-power reference under the same recipe and seed. Repeat promising comparisons with seeds 43 and 44 before claiming improvement.

Record subset and full validation separately, local second-moment cosine, tensor and worker consensus error energy, maximum-coordinate energy fraction, correction norm relative to the Adam step, and worker-mean drift. Instrument before and after mixing and after optimizer updates. Recheck directional curvature of the actual fresh disagreement and its shaped counterpart on identical checkpoint and batch inputs, including an FP32 check. Success requires stable training and replicated validation gains; agreement of moments or improved one-step curvature alone does not satisfy it.

7 Presentation speaking notes

These notes supplement the short slides without adding to the audience-facing deck.

Slide	Interpretation to emphasize
1	The 20.4M count is per local model, and 408M is the global token budget. “Epoch” is a reporting interval.
2	Packed workers share hardware, while parameters, data assignments, and Adam histories remain local. Evaluation averages parameters.
3	The active-interval LR normalization makes weakening begin smoothly; full mixing means the base topology’s mixing operation.
4	Noise becomes larger relative to the mean gradient by the final checkpoint; absolute norms and cross-run batch sizes tell different stories.
5	The loss gap is small and single-seed; it supports no observed advantage for this recipe.
6	Alignment is a pooled directional curvature, not a cosine or a normalized probability.
7	Large diagnostic gains are measured at unchanged baseline checkpoints; this motivates an experiment rather than proving its outcome.
8	Second moments agree much more than first moments or Adam directions; high cosine does not imply exact equality.
9	The algorithm transforms existing consensus error direction every step, using previous-step moments; norm preservation is per worker.
10	Topology differs from the first comparison. The matched ordinary-AC control is missing; full mixing is the strongest saved reference.
11	A stable total norm can conceal almost all energy moving into one coordinate. Frozen replays demonstrate mechanisms rather than new training results.
12	The next method limits shared shaping to newly generated update disagreement. Its training value was still to be tested at the review stage.

8 Source map and reproducibility

The package archives all six requested original PDFs byte for byte. The appendix displays five unchanged originals and redraws the review figure with the requested consensus-error terminology, preserving its measurements. The slides redraw selected curves from saved CSV/JSONL measurements without smoothing. The package build requires only the included source/data/assets, Python plotting dependencies, and Typst; it does not train a model or recompute Hessian products. Full input hashes, selected series, and transformations are recorded in the package provenance; original PDF hashes also appear in `assets/originals/sources.json`. `historical-source-provenance.json` records hashes of the executed source files used to check the equations.

S1 · Synchronous run: resolved recipe, result, metrics, and alignment analysis
`runs/20m-0-0022/`

S2 · Plain exponential-topology packed run
`runs/packed-4-20m-0-0048/`

S3 · Ordinary exponential-topology adaptive-consensus run and original alignment
`runs/packed-ac-4-20m-0-0048/`

S4 · Seven-family scaled-direction analysis on ordinary AC checkpoints
`runs/packed-ac-4-20m-0-0048/analysis/scaled-alignment/`

S5 · Full-state optimizer similarity estimator and measurements
`runs/packed-ac-4-20m-0-0048/analysis/similarity/full/`

S6 · Run review, checkpoint measurements, and frozen-state replays
`runs/consensus-scaling-review-2026-09-11/`

S6 · principal run · Complete-topology, alpha=0.5 training and subsequent diagnostics
`runs/packed-ac-full-scale-4-20m-0-0048/`

S7 · Historical ordinary-AC source: model, analysis/core.py, train.py, packed.py
`runs/packed-ac-4-20m-0-0048/analysis/alignment/source/tiny_llm/`

S8 · Historical scaled source: train.py, packed.py, and packed_optimizer.py
`runs/packed-ac-full-scale-4-20m-0-0048/analysis/alignment/source/tiny_llm/`

S4 preserves its numerical policy in `README.md` and the executed seven-family analysis in `source-v2/scaled_analysis.py`. S5′s estimator is `optimizer_similarity.py` in its parent directory. The scaled run also preserves a source snapshot under `analysis/scaled-alignment/source/`; the training equations here are cross-checked against its `analysis/alignment/source/` snapshot. Source identifiers match the slide footer references.

The review's `run_comparison.csv` gives all twelve final-run entries. `checkpoint_summary.csv`, `parameter_breakdown.csv`, and `largest_error_coordinates.csv` establish concentration and layer identities. `counterfactual_mixing.csv`, `fresh_update_counterfactual.csv`, and `algebra_checks.json` record the replays and algebra checks. `provenance.json` inside the review records checkpoint paths, sizes, modification times, recipes, and saved positions; the review checks epoch/step identities when reading. The source snapshots preserve the actual executed files despite uncommitted development between historical runs.

The run abbreviations in the table map directly to S2/S3 and the following directories under `runs/`: `packed-ac-exclude-4-20m-0-0048`, `packed-ac-exclude-full-4-20m-0-0048`, `packed-full-4-20m-0-0048`, and S6 with suffixes `-t2` through `-t7`. No external theoretical result is used as a guarantee for the tested or proposed disagreement transformations.

Original figure A · Synchronous gradient norms

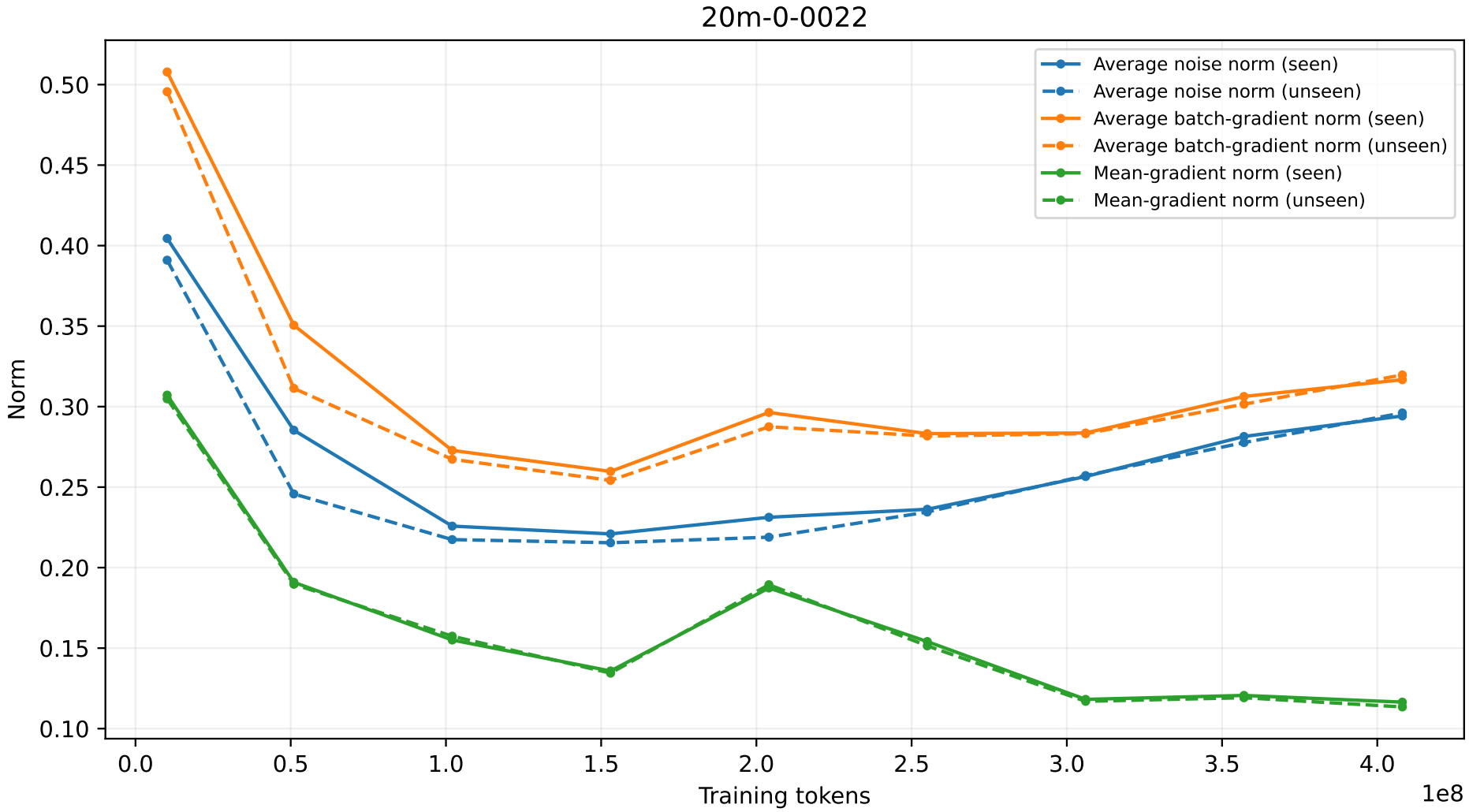

Figure 1: Source S1, `analysis/alignment/gradient-norms.pdf`. Both seen and unseen cases are preserved. Gradient noise is measured from 32 diagnostic batches at the checkpoint, with a full-epoch mean gradient as reference. The presentation selects unseen curves and reports the ratio of average noise norm to mean-gradient norm.

Original figure B · Packed gradient norms

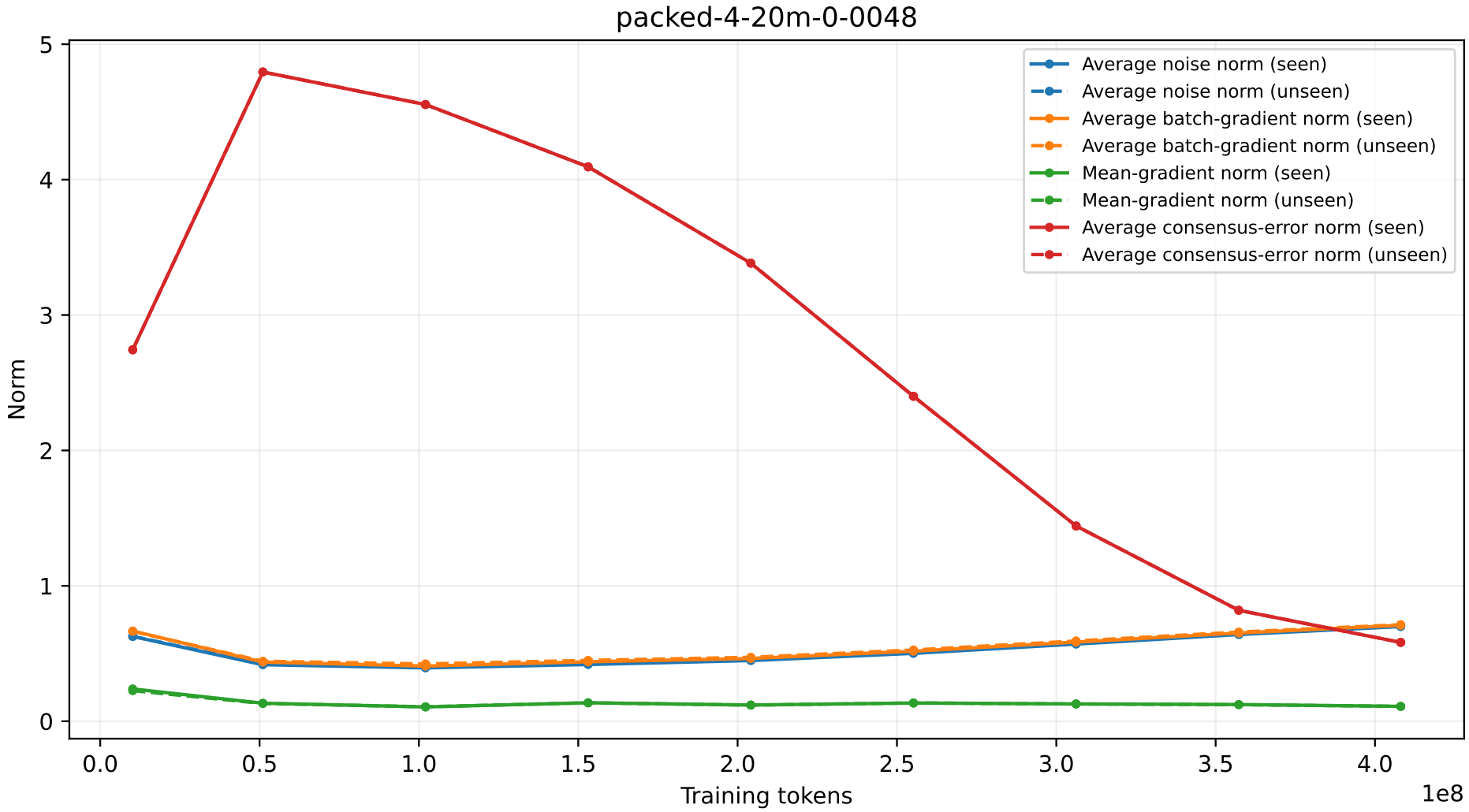

Figure 2: Source S2, `analysis/alignment/gradient-norms.pdf`. Ordinary four-worker exponential-topology mixing. The noise batch is the worker-local batch of 8,192 tokens; its absolute scale should not be directly compared with the synchronous batch of 32,768 tokens without this qualification.

Original figure C · Ordinary AC alignment

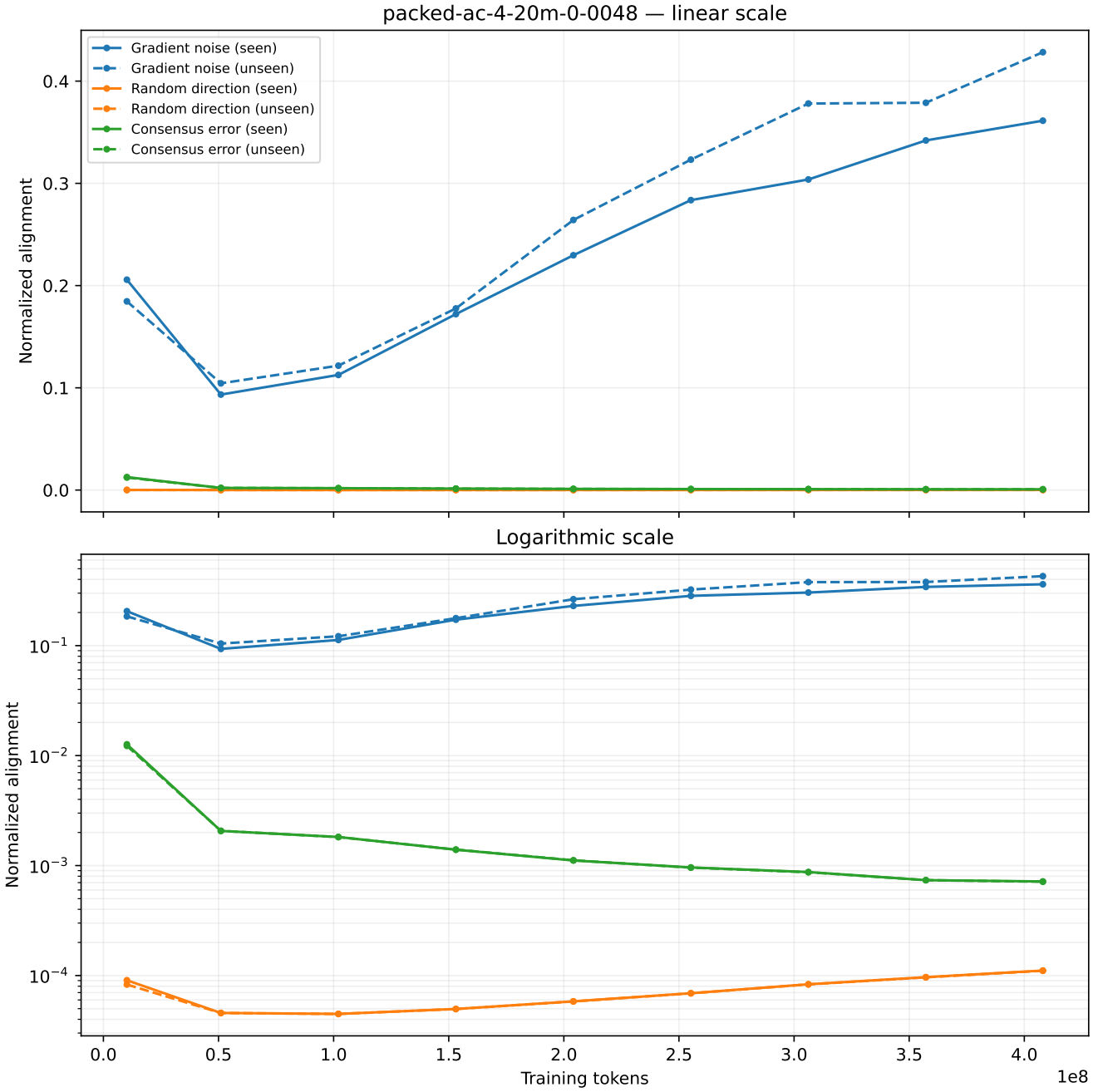

Figure 3: Source S3, `analysis/alignment/normalized-alignment.pdf`. The original linear and logarithmic panels retain seen/unseen curves, gradient-noise directions, consensus errors, and the random reference. “Normalized alignment” denotes the pooled quadratic form defined in Section 3.

Original figure D · Moment-weighted alignment

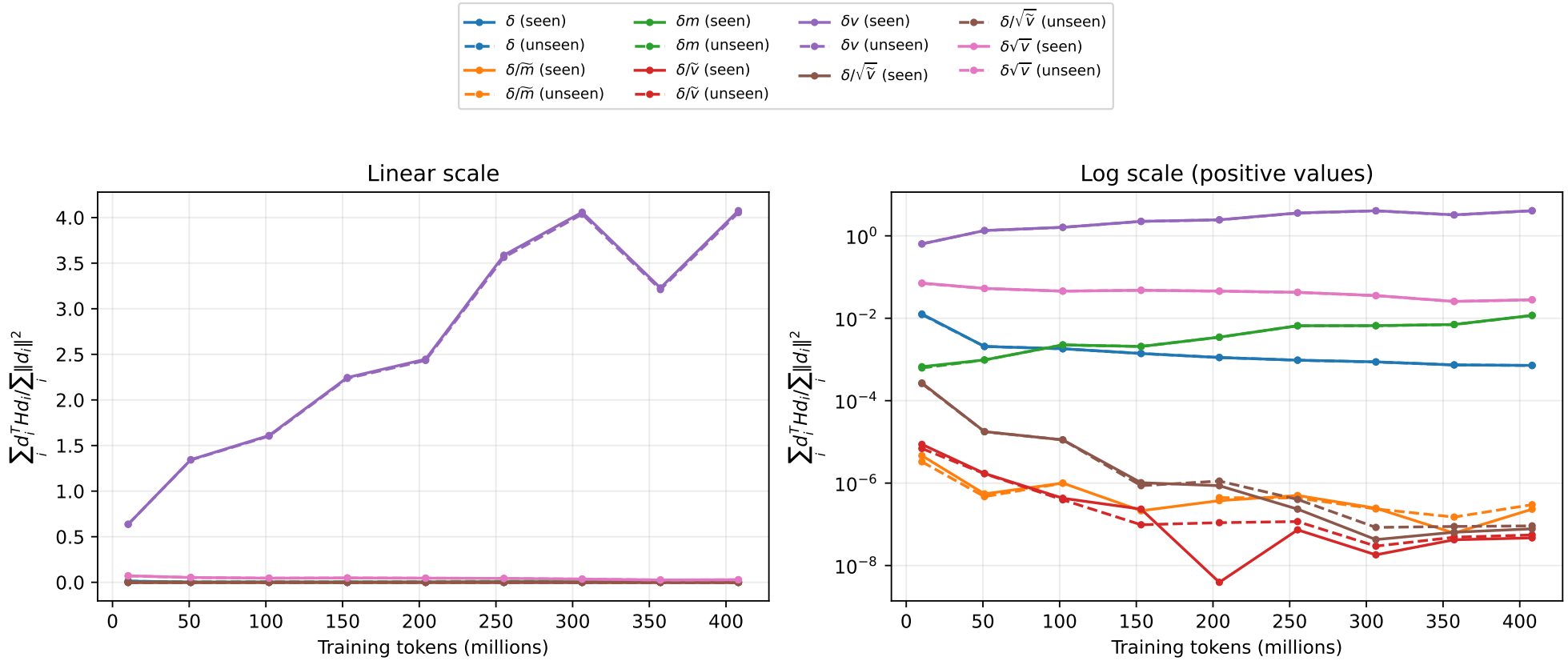

Figure 4: Source S4, `normalized-alignment.pdf`. The current seven-family extension includes multiplication and division by first moments, second moments, and their square roots. This diagnostic transforms directions on saved ordinary-AC checkpoints; it does not rerun training.

Original figure E · Full-state worker similarity

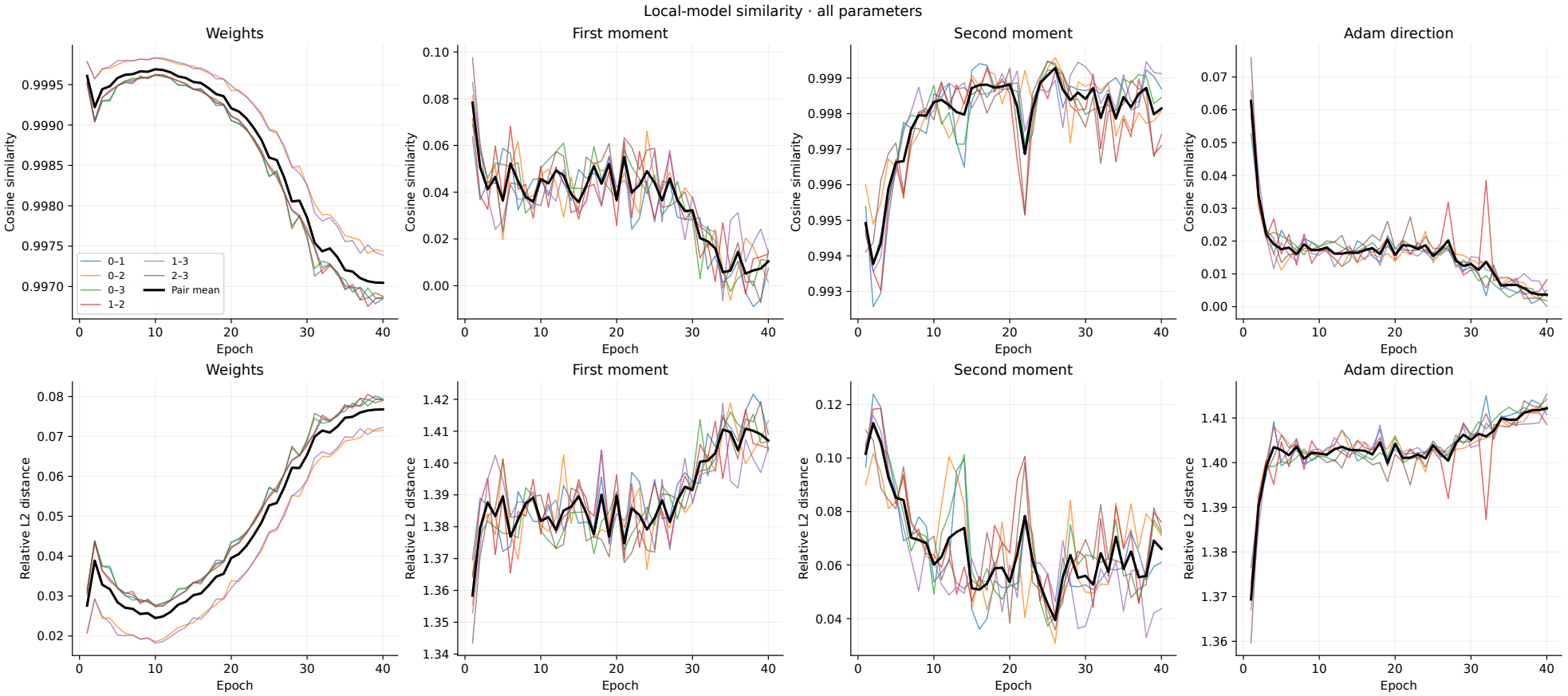

Figure 5: Source S5, `similarity-curves.pdf`. All unique parameters, including tied embedding/output weights counted once, across 40 epochs and six unordered worker pairs. Colored curves show the six worker pairs; black curves show their mean. These are not independent training seeds. The full-state view establishes the pronounced difference between second moments and Adam directions.

Review figure F · Scaling review

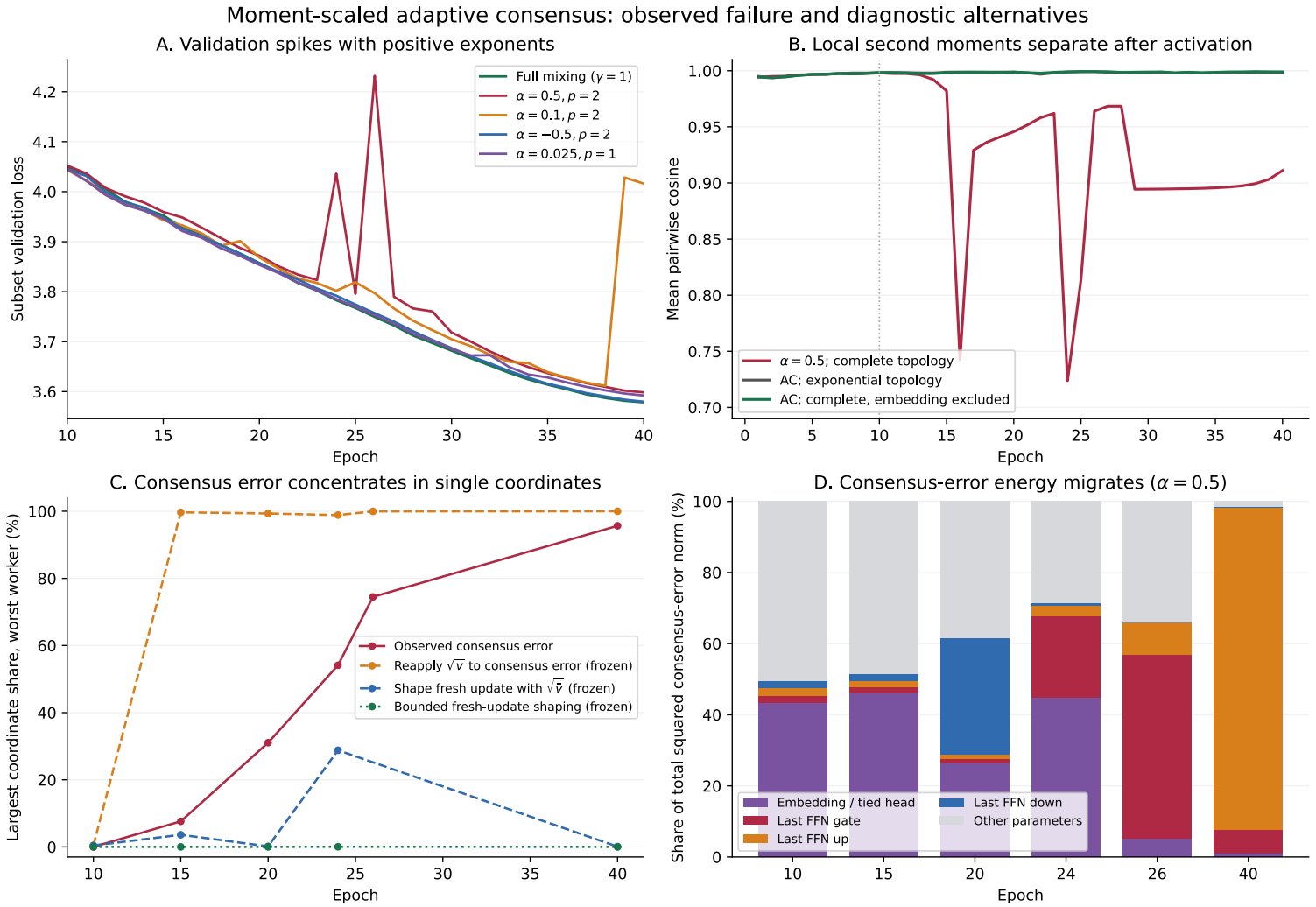

Figure 6: Source S6, `diagnosis.pdf`. Redrawn with consensus-error terminology; the same measurements combine observed training, checkpoint concentration, and frozen-state alternatives. These categories are interpreted separately in Sections 4–6. The full-mixing baseline has the best observed complete-topology validation result; the proposed fresh-update method has no training result in this review.