Untangling Component Imbalance in Hybrid Linear Attention Conversion Methods

HedgeCATs, SSD, and the meaning of successful conversion

Research paper discussion

arXiv:2510.05901v2 | 15 October 2026

Paper authors: Martin Benfeghoul, Teresa Delgado, Adnan Oomerjee, Haitham Bou-Ammar, Jun Wang, and Zafeirios Fountas.

Huawei Noah's Ark Lab and University College London. Paper revised 10 October 2025.

A high hybrid score can hide a failed linear branch

The central question

After converting a pretrained Transformer, does linear attention actually contribute to the model's performance?

Mistral-7B	Hybrid	SWA only	LA only
Six-task average (%)	65.48	65.56	34.78

- Removing LA slightly improves the reported average.
- Removing all attention gives 34.40: close to LA only.
- Overall recovery and component usefulness need separate tests.

Paper, Sec. 3.2 and Table 1, PDF pp. 4–5. Mistral results use the authors' implementation.

LA compresses history; SWA retains local detail

$$\mathbf{S}_t = \sum_{i \le t} \phi(\mathbf{k}_i)\mathbf{v}_i^\top, \qquad \mathbf{z}_t = \sum_{i \le t} \phi(\mathbf{k}_i), \qquad \mathbf{o}_t^{\mathrm{LA}} = \frac{\phi(\mathbf{q}_t)^\top \mathbf{S}_t}{\phi(\mathbf{q}_t)^\top \mathbf{z}_t}$$

- $\phi : \mathbb{R}^d \to \mathbb{R}^{d'}$; $\mathbf{S}_t \in \mathbb{R}^{d' \times d_v}$ and $\mathbf{z}_t \in \mathbb{R}^{d'}$.
- Each new token adds one outer product and one vector.

The hybrid studied here

Recent w tokens $\longrightarrow$ SWA ; older tokens $\longrightarrow$ LA state

Hybrid output: $\mathbf{o}_t = \frac{1}{2}\mathbf{o}_t^{\mathrm{SWA}} + \frac{1}{2}\mathbf{o}_t^{\mathrm{LA}}$.

Fixed w and feature size give linear sequence scaling. Equal coefficients do not imply equal functional contributions.

Paper, Eqs. (2)–(3), Sec. 3.1, PDF pp. 2–4. Causal form above is an explanatory rewrite.

Feature capacity and transfer objective are separate choices

$$\phi(\mathbf{x}) = [\,\sigma(\mathbf{W}_\phi^\top \mathbf{x} + \mathbf{b}) \oplus \sigma(-\mathbf{W}_\phi^\top \mathbf{x} - \mathbf{b})\,]$$

Projection	$\mathbf{W}_\phi$ shape	Output dimension
Square	$h_d \times h_d$	$d' = 2h_d$
Rectangular	$h_d \times (h_d/2)$	$d' = h_d$

- Positive and negative projections preserve both directional responses.
- Default activation: softmax over feature coordinates.
- RoPE is applied before ϕ; smaller d' also reduces state size.

Paper, Eq. (4), Secs. 3.1 and 3.4.2, PDF pp. 3–4 and 6. h_d: query/key head dimension.

The hybrid loss can let SWA carry the approximation

With teacher weights A^{soft} and output O^{soft}:

$$\mathcal{L}_{\text{weights}} = -\sum_{t,i\le t} A_{ti}^{\text{soft}} \log A_{ti}^{\phi} \qquad \text{LA weights}$$

$$\mathcal{L}_{\text{LA-out}} = \|O^{\text{soft}} - O^{\text{LA}}\|_F^2 \qquad \text{LA output}$$

$$\mathcal{L}_{\text{hybrid-out}} = \|O^{\text{soft}} - [gO^{\text{SWA}} + (1-g)O^{\text{LA}}]\|_F^2 \qquad \text{joint output}$$

- The first two objectives require LA to perform the approximation.
- The third supervises only the combined result.
- Weight transfer materializes pairwise weights; a linear student output still does not remove the softmax teacher's quadratic work.

Paper, Sec. 3.4.1, PDF p. 6. Equations organize the stated objectives; averaging factors omitted.

The experiments isolate conversion under a small data budget

Models and training

- Mistral-7B, Llama3-8B, Llama3.1-8B: instruction-tuned models.
- FineWeb-Edu; 1,024-token samples; about 25M tokens/epoch.
- Transfer: 1 epoch.
- LoRA: 1–5 epochs; $r = 8$, $\alpha = 16$ on Q, K, V, O projections.

Evaluation

- Zero-shot PIQA, ARC-E, ARC-C, HellaSwag, WinoGrande, MMLU.
- ARC-C and HellaSwag: normalized accuracy.
- AVG: equal mean of six task accuracies.
- Recovery: converted AVG divided by original AVG.

Paper, Sec. 3.1, PDF p. 4. Default $g = 0.5$, SWA window 64; LA excludes the local window.

SWA carries most of the LoLCATs-style hybrid performance

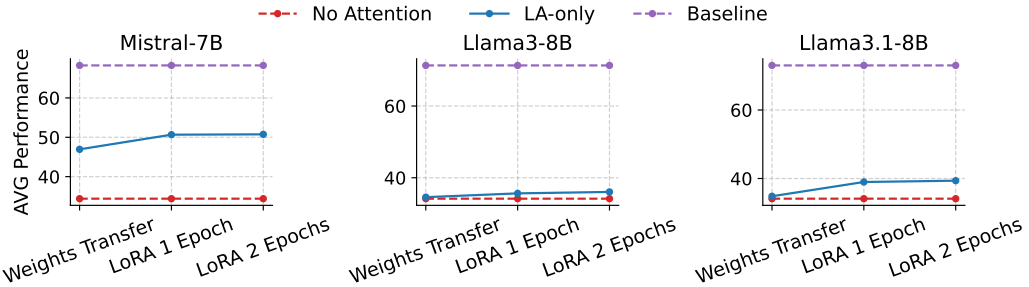
Active attention	Mistral-7B	Llama3-8B	Llama3.1-8B*
Original model	68.26	71.32	73.01
Hybrid	65.48	67.55	71.08
SWA only	65.56	66.08	71.45
LA only	34.78	33.55	33.96
No attention	34.40	34.17	33.93
Hybrid − SWA only (pp)	−0.08	+1.47	−0.37

- LA alone is close to the no-attention control in all three cases.
- Llama3 retains a positive LA contribution; complete bypass is not universal.

Paper, Table 1, PDF p. 5. Accuracy in %; differences calculated from reported AVG.

*Official LoLCATs checkpoints and code; the other two columns use the authors' implementation.

Pure LA works better for Mistral than for the Llama models



- Mistral: 46.94 after weight transfer $\rightarrow$ 50.75 after 2 LoRA epochs.
- Llama3: 34.59 $\rightarrow$ 36.05; Llama3.1: 34.82 $\rightarrow$ 39.36.
- LoRA is not intrinsically incompatible with LA, but conversion difficulty depends strongly on the starting model.

Paper, Fig. 1 and Sec. 3.3, PDF p. 5; Table 6, PDF p. 15. Original author figure.

A larger map does not rescue hybrid-output transfer

Mistral-7B; evaluate LA only after one transfer epoch, without LoRA.

Transfer target	$d' = 2h_d$	$d' = h_d$
Attention weights (CE)	46.94	43.25
LA outputs (MSE)	47.23	42.60
Hybrid outputs (MSE)	35.41	34.59
No attention	34.40	34.40

- At $d' = 2h_d$, pure-output transfer beats hybrid transfer by 11.82 pp.
- Reducing capacity hurts both successful transfer objectives.
- Output MSE can work: the hybrid target is the critical distinction.

Paper, Tables 2–3 and Secs. 3.4.1–3.4.2, PDF p. 6. AVG in %; pp = percentage points.

Exponential-family feature maps transfer more successfully

Mistral-7B; square projection; one weight-transfer epoch; LA-only evaluation.

Activation	AVG (%)	Above no attention (pp)
Softmax	46.94	12.54
Exponential	46.11	11.71
ReLU	36.07	1.67
$1 + \mathrm{ELU}$	35.12	0.72
None	34.81	0.41

Softmax and exponential activations are close in this protocol; the alternatives are much closer to the no-attention score of 34.40.

Paper, Table 4 and Sec. 3.4.3, PDF p. 7. Differences calculated from reported AVG.

Three interventions change when SWA can take over

1. Inference-time hybrid

Train LA first; add SWA only at inference.
No additional hybrid training is needed.

2. HedgeCATs

LA weight transfer $\longrightarrow$ brief hybrid LoRA.
Early stopping limits the drift back toward SWA dependence.

3. Scheduled Sliding-window Dropout (SSD)

Randomly suppress SWA during hybrid LoRA.
The linear path must operate when the local branch is unavailable.

Paper, Secs. 4.1–4.3, PDF pp. 7–8. "Zero-shot hybrid" still includes prior LA training.

Adding SWA at inference reveals useful complementarity

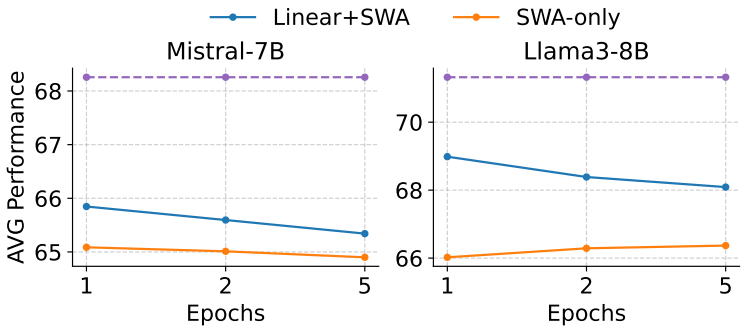
After weight transfer, with no LoRA; default window 64.

Model	LA only	Hybrid	Overlap	SWA only
Mistral-7B	46.94	66.02	63.34	61.66
Llama3-8B	34.59	67.91	51.41	61.83
Llama3.1-8B	34.82	68.71	56.04	62.57

- Removing LA costs 4.36, 6.08, and 6.14 pp, respectively.
- Weak standalone LA can still complement SWA.
- Overlap hurts at fixed $g = 0.5$; adaptive mixing is suggested as a remedy.
- Llama3.1 recovery is 94.11%, below the paper's broad 95% claim.

Paper, Table 8, PDF p. 16; Sec. 4.1, PDF pp. 7–8. AVG in %; recovery as reported for this row.

HedgeCATs needs a short hybrid fine-tuning schedule



- Best tested epoch: 1. AVG = 65.85 for Mistral, 68.99 for Llama3.
- LA ablation gap shrinks from 0.76 to 0.44 pp for Mistral, and from 2.97 to 1.72 pp for Llama3 (epochs $1 \rightarrow 5$).

Paper, Fig. 3, PDF p. 8; Table 9, PDF p. 17. Original author figure; dashed lines are original models.

SSD removes the local shortcut during training

At epoch k

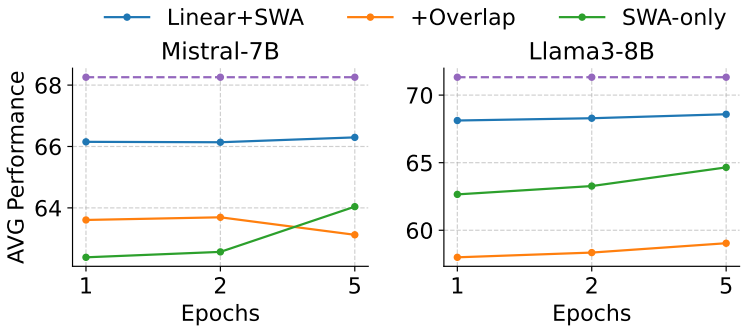
Drop the SWA branch with probability p_k and use window w_k. Hold the last schedule value after the schedule ends.

Experiment	Drop probability p_k	Window w_k
Dropout schedule	$0.9 \to 0.75 \to 0.5$	32
Window schedule	0.5	$4 \to 8 \to 16 \to 32 \to 64$
Fixed setting	0.5	16

- Heavy early dropout forces more training through LA.
- Mask granularity and branch rescaling are not fully specified; standard inverted-dropout scaling should not be assumed.

Paper, Sec. 4.3, PDF p. 8. Tables 10–12, PDF pp. 17–18.

Scheduled dropout sustains accuracy and a positive LA gap



- At 5 epochs: Mistral 66.30 vs. 64.04 SWA-only; Llama3 68.58 vs. 64.65.
- LA adds 2.26 and 3.93 pp, but the gaps still shrink with training.
- Setting: $p = 0.9 \rightarrow 0.75 \rightarrow 0.5$, fixed window 32.

Paper, Fig. 4a, PDF p. 9; Table 10, PDF p. 17. Original author figure; AVG accuracy in %.

Growing the window does not improve this conversion

Model	SSD configuration	1 epoch	2 epochs	5 epochs
Mistral	Scheduled p, $w = 32$	66.15	66.14	66.30
	Fixed $p = 0.5$, $w = 16$	64.34	65.99	64.97
	Fixed p, growing w	64.03	62.43	61.42
Llama3	Scheduled p, $w = 32$	68.12	68.29	68.58
	Fixed $p = 0.5$, $w = 16$	68.18	68.03	68.45
	Fixed p, growing w	67.85	67.51	67.45

- Fixed dropout already works reasonably well.
- The experiments also change window size: they do not isolate the benefit of scheduling dropout at a matched window.

Paper, Tables 10–12, PDF pp. 17–18. Hybrid AVG in %. Growing w: 4, 8, 16, 32, 64.

Average recovery hides a persistent MMLU deficit

Llama3-8B; SSD dropout schedule, window 32, 5 LoRA epochs.

Task	Original model	SSD hybrid	SWA only
PIQA	78.13	78.13	77.48
ARC-Easy	81.69	80.72	80.43
ARC-Challenge	56.66	56.23	55.12
HellaSwag	75.94	72.84	56.11
WinoGrande	71.67	72.61	72.61
MMLU	63.85	50.97	46.16
AVG	71.32	68.58	64.65

HellaSwag gains 16.73 pp from LA; MMLU still trails the original by 12.88 pp.

Paper, Table 10, PDF p. 17. Accuracy in %; ARC-C and HellaSwag use normalized accuracy.

Accuracy and LA dependence favor different choices

Method	Mistral-7B AVG	Mistral-7B Δ_{LA}	Llama3-8B AVG	Llama3-8B Δ_{LA}
LoLCATs-style diagnostic	65.48	-0.08	67.55	$+1.47$
Inference hybrid; no LoRA	66.02	$+4.36$	67.91	$+6.08$
HedgeCATs; 1 epoch	65.85	$+0.76$	68.99	$+2.97$
SSD; 5 epochs	66.30	$+2.26$	68.58	$+3.93$

- $\Delta_{\mathrm{LA}} = \mathrm{AVG}_{\mathrm{hybrid}} - \mathrm{AVG}_{\mathrm{SWA-only}}$, in percentage points.
- Inference hybrid: no joint training; largest gaps in these rows.
- HedgeCATs: highest selected Llama3 score with brief LoRA.
- SSD: longer adaptation, smaller window (32 vs. 64).

Paper, Tables 1, 8–10, PDF pp. 5, 16–17. Selected settings, not a matched-budget ranking.

The evidence supports useful components in specific settings

- **Functional dependence:** removing LA tests its usefulness in a trained hybrid. A 50/50 coefficient does not establish equal contribution.
- **Context and efficiency:** six zero-shot tasks establish the reported accuracy effects; long-context retrieval, latency, and memory savings are not measured here.
- **Recovery depends on the setting:** Llama3.1 inference hybrid reaches 94.11%; the growing-window SSD can be substantially worse.
- **Reporting conflicts matter:** some recovery ratios and prose trends disagree with the appendix. Use the explicit AVG rows for comparisons.

Paper, Secs. 4–5, PDF pp. 7–10; Tables 8–12, PDF pp. 16–18. Specific conflicts in backup slide 26.

Conversion needs accuracy and useful components

1. **Diagnose the complete model.**
   Report hybrid, SWA-only, LA-only, and no-attention controls.
2. **Give LA an independent learning signal.**
   Weight transfer and pure-LA output transfer both outperform hybrid-output transfer in the main diagnostic.
3. **Control later competition between branches.**
   Delayed hybridization, early stopping, and branch dropout offer different accuracy, attribution, and budget trade-offs.

A concrete next experiment

Combine component ablations with distance-controlled retrieval and measured inference costs at matched training budgets.

Synthesis of the paper's results; the proposed next experiment is this reading's analysis.

Paper and presentation scope

Untangling Component Imbalance in Hybrid Linear Attention Conversion Methods

Martin Benfeghoul, Teresa Delgado, Adnan Oomerjee, Haitham Bou-Ammar, Jun Wang, and Zafeirios Fountas.

arXiv:2510.05901v2, revised 10 October 2025.

- All experimental values come from this version of the paper.
- Figures retain the authors' original plots and legends.
- Tables select relevant rows; derived gaps are labeled in percentage points.
- Backup slides: mixing weights, additional model results, complete inference curves, fixed SSD, and reporting conflicts.

Research discussion prepared 15 October 2026. The presentation author is not a paper author.

Backup: a fixed mixing coefficient is not a softmax partition

Partition the history into a recent window and older positions. If both branches were exact softmax:

$$Z_{\text{near}} = \sum_{i \in \mathcal{W}_t} e^{\mathbf{q}_t^\top \mathbf{k}_i / \sqrt{d}}, \qquad Z_{\text{far}} = \sum_{i \notin \mathcal{W}_t,\ i \le t} e^{\mathbf{q}_t^\top \mathbf{k}_i / \sqrt{d}}$$

$$\mathbf{o}_t^{\text{full}} = \alpha_t \mathbf{o}_t^{\text{near}} + (1 - \alpha_t)\mathbf{o}_t^{\text{far}}, \qquad \alpha_t = \frac{Z_{\text{near}}}{Z_{\text{near}} + Z_{\text{far}}}.$$

- α_t depends on the query and history; it is generally not 0.5.
- An overlapping LA branch sends recent tokens through both paths.
- This motivates adaptive mixing, but does not establish its benefit or eliminate LA approximation error.

Explanatory derivation from softmax normalization; related to the authors' discussion in Sec. 4.1, PDF p. 7.

Backup: pure-output transfer also helps the Llama models

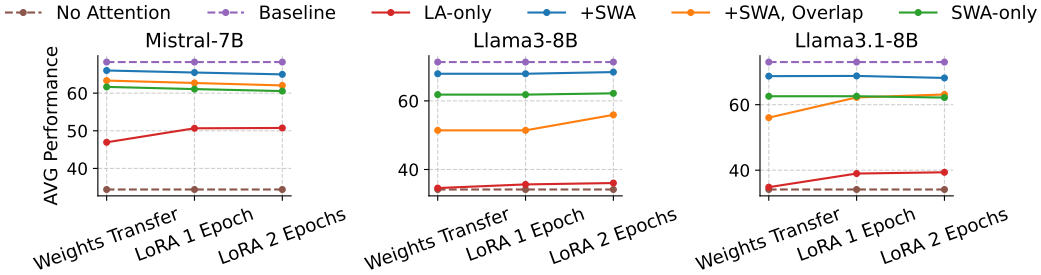
After one transfer epoch; square projection; no LoRA; LA-only evaluation.

Transfer target	Mistral-7B	Llama3-8B	Llama3.1-8B
Weights	46.94	34.59	34.82
LA outputs	47.23	38.22	35.45
Hybrid outputs	35.41	33.86	33.87
No attention	34.40	34.17	34.11

- Llama3 benefits more from pure-output than from weight transfer.
- LA conversion remains far below the original Llama model scores.
- HedgeCATs uses weight transfer, but the evidence does not make that objective uniformly best.

Paper, Table 7, PDF p. 15. AVG accuracy in %.

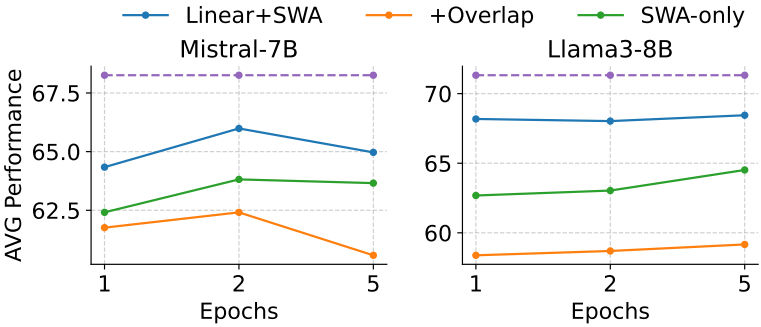
Backup: no-LoRA hybridization is not uniformly the best row



- Mistral hybrid: 66.02 → 65.48 → 64.99 as pure-LA LoRA proceeds.
- Table 8 lists Llama3 as 67.91 → 67.91 → 68.40, and Llama3.1 as 68.71 → 68.78 → 68.16.
- These rows do not support a universal "best without fine-tuning" claim; the Llama3 repeated rows and ratios also warrant caution.

Paper, Fig. 2, PDF p. 8; Table 8, PDF p. 16. Original author figure.

Backup: fixed dropout already preserves a useful linear path



- Fixed $p = 0.5$, $w = 16$: Mistral peaks at epoch 2 (65.99).
- Llama3 reaches 68.45 at epoch 5 vs. 64.52 for SWA only.
- Scheduled dropout should be compared at the same window before attributing an improvement specifically to scheduling.

Paper, Fig. 4c, PDF p. 9; Table 11, PDF p. 18. Original author figure.

Backup: several reporting conflicts change the interpretation

- **Inference hybrid, Table 8:**
  Llama3, 2 epochs: AVG 68.40 / original 71.32 implies about 95.91% recovery; the table prints 95.27%.
- **SSD, Table 10:**
  Llama3, 5 epochs: AVG 68.58 / 71.32 implies about 96.16%; the table prints 95.97%.
- **Window scheduling, Table 12 and Fig. 4b:**
  Llama3 hybrid falls 67.85 $\rightarrow$ 67.51 $\rightarrow$ 67.45. Sec. 4.3 describes a mild late recovery.
- These slides retain the reported AVG values and flag the conflicts; they do not infer corrected experimental measurements.

Paper, Sec. 4.3, PDF p. 8; Fig. 4b, PDF p. 9; Tables 8, 10, 12, PDF pp. 16–18.