# Dual-State Linear Attention and Adaptive Layer Conversion

## A reading of "On-the-Fly Adaptive Distillation of Transformer to Dual-State Linear Attention"

Paper: Yeonju Ro, Zhenyu Zhang, Souvik Kundu,
Zhangyang Wang, Aditya Akella

arXiv:2506.09316v3 | June 17, 2025
Reading presentation | October 8, 2026

Two memory timescales; a runtime choice of how many layers to convert.

Paper PDF adds the subtitle "for Long-Context LLM Serving." https://arxiv.org/abs/2506.09316v3

# Two pressures motivate a flexible attention model

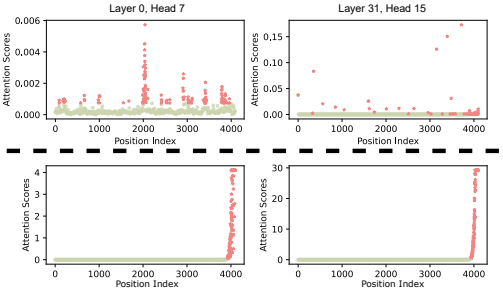



Fig. 1, PDF p. 3. Transformer above; distilled GLA below. Axes have different scales.

**Memory quality**

- A single gated state emphasizes recent tokens in the displayed heads.
- Earlier information passes through more forget gates.

**Serving resources**

- Long sessions accumulate KV cache.
- Bursts and session affinity create GPU hotspots.

Sec. 3, PDF pp. 2–3. Fig. 1: 16 C4 samples, 4096 tokens; GLA distilled with 1B tokens.

# DSLA writes the same token into two gated states

$$\mathbf{S}_t^1 = \mathbf{G}_t^1 \odot \mathbf{S}_{t-1}^1 + \mathbf{k}_t^\top \mathbf{v}_t,$$
$$\mathbf{S}_t^2 = \mathbf{G}_t^2 \odot \mathbf{S}_{t-1}^2 + \mathbf{k}_t^\top \mathbf{v}_t,$$
$$\mathbf{o}_t = \mathbf{q}_t(\gamma \mathbf{S}_t^1 + (1-\gamma)\mathbf{S}_t^2).$$

- Per head: $\mathbf{q}, \mathbf{k}, \mathbf{v} \in \mathbb{R}^{1\times d}$; states and gates are $d \times d$.
- Shared write, different forgetting: history and recency emerge through training.
- One learned $\gamma$ per layer mixes the two readouts.

Sec. 4.1, Eqs. (4)–(5), PDF p. 4. Eq. (4) omits old-state branch superscripts; the equations above make the independent-state interpretation explicit.

# Two gates create complementary retention timescales

$$\mathbf{S}_T^a = \sum_{t=1}^{T} \mathbf{P}_{t\to T}^a \odot (\mathbf{k}_t^\top \mathbf{v}_t), \qquad \mathbf{P}_{t\to T}^a = \bigodot_{\tau=t+1}^{T} \mathbf{G}_\tau^a.$$

## Scalar teaching example: the retention part of the weight

$$w(\Delta) = \tfrac{1}{2}(0.999)^\Delta + \tfrac{1}{2}(0.9)^\Delta$$

History half-life: $\sim 693$ tokens. Recency half-life: $\sim 6.6$ tokens.

- Both states receive every token; there is no explicit context partition.
- Content matching through query/key still determines what is read.

Teaching derivation from the recurrence; zero initial states. Related to Eqs. (6)–(8), PDF p. 4. Example gates are illustrative, not measured parameters.

# Teacher alignment and gate diversity play different roles

$$\mathcal{L} = \underbrace{\mathcal{D}_{KL}(\mathbf{o}_t \| \mathbf{o}_t^{(\mathrm{gt})})}_{\text{match teacher output}} + \lambda \underbrace{\mathrm{sim}(\mathbf{G}_t^1, \mathbf{G}_t^2)}_{\text{cosine similarity}}.$$

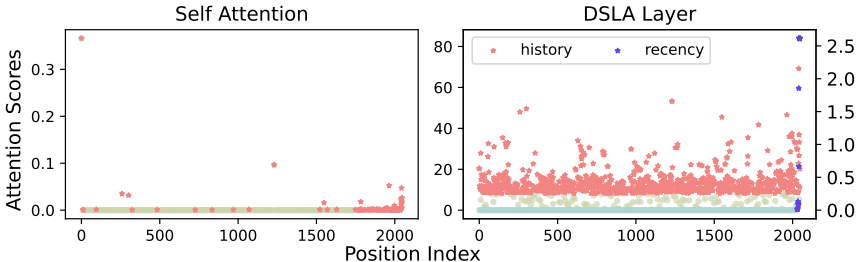
- Minimizing similarity discourages the two gates from collapsing together.
- Q/K/V/O projections reuse teacher weights; gate initialization differs by branch.
- The paper does not specify the probability normalization needed for KL on attention outputs, or the value of $\lambda$.

## Reading the result

The objective explains the intended division of labor; a numerical implementation needs additional choices.

Sec. 4.2, Eqs. (9)–(10), PDF p. 4; initialization in App. B, p. 14.

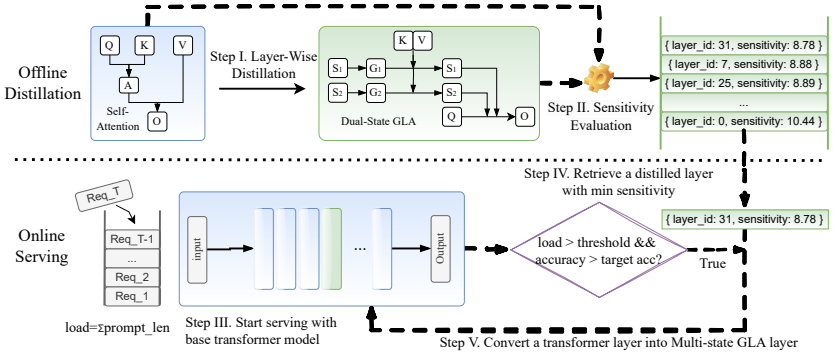
# The visualized states focus on different token regions




- Left: teacher self-attention. Right: history in red, recency in blue.
- History covers earlier positions; recency concentrates near the end.
- The two DSLA series use different vertical scales: compare location patterns.

Sec. 5.5, Fig. 5, PDF p. 8. Scores are not a shared softmax probability distribution.

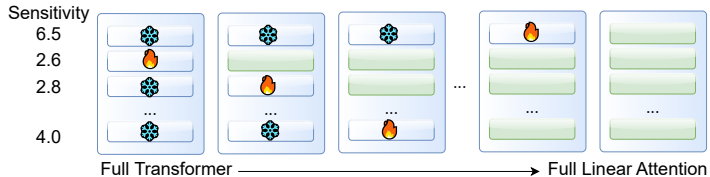
# Offline distillation prepares the choices used online




**Offline:** prepare replacements. **Online:** choose layers under load.

App. A, Fig. 8, PDF p. 14; Sec. 4.3, pp. 4–5. Runtime conversion uses trained substitutes.

# Chained tuning follows the deployment conversion order




1. Rank layers from least to most sensitive.
2. Replace one layer; train it while freezing all other layers.
3. Keep that replacement, then train the next layer in the updated model.

$$M_k : \quad \{\pi_1, \ldots, \pi_k\} \text{ use DSLA.}$$

Sec. 4.4, Fig. 3, PDF p. 5. This directly trains nested prefixes of a fixed order; it does not enumerate arbitrary subsets of layers.

# Attention entropy provides a sensitivity heuristic

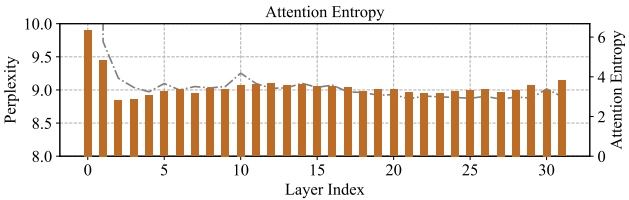



Fig. 7, PDF p. 9. Bars: WikiText-2 PPL after one layer is replaced. Line: entropy.

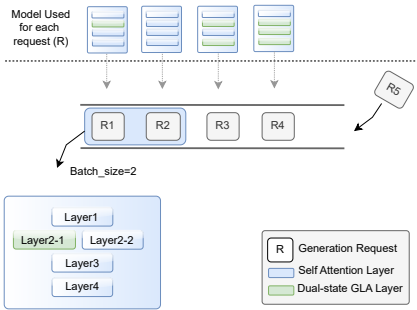
- Prefer layers with smaller validation impact.
- Concentrated attention motivates a cheaper entropy proxy.
- This is an empirical ranking heuristic.

## A sign matters for the sorting direction

Standard entropy: $H = -\sum_i A_{T,i} \log A_{T,i}$. Eq. (11) omits the minus sign, although the prose uses standard entropy intuition.

Sec. 4.3, Eq. (11), PDF p. 5; App. F, pp. 16–17.

# Mixed requests share layers, then split at a conversion




- Requests can use different conversion depths.
- Shared layers batch together.
- Divergent layers require separate DSLA and self-attention sub-batches.
- The authors argue KV savings outweigh re-batching overhead.

Mid-sequence switching also requires a valid history state; its construction cost is not fully specified.

Sec. 4.5, PDF pp. 5–6; App. D, pp. 15–16. Fig. 11: left diagram cropped, full legend retained.

# Evaluation separates model quality from serving latency

| Setting | Configuration / comparison role |
|---|---|
| Main conversion | Llama2-7B teacher; 1.6B SlimPajama tokens; up to 4K training length; 25% or 50% layers converted |
| Quality baselines | Teacher, GLA, Mamba, RetNet (selected short tasks), and static hybrid Zamba |
| Quality metrics | LongBench scores, PPL, ROUGE, seven short-task scores |
| Latency hardware | A100 80GB for Fig. 4; A6000 for the end-to-end replay in Table 4 |

The 1.6B tokens are additional distillation data; teacher pretraining is inherited.

Sec. 5, PDF pp. 6–9; App. B, p. 14. Cross-model training budgets and task protocols are not all matched.

# Partial conversion improves QA but worsens some PPL

| Model | WikiText PPL ↓ | Lambada PPL ↓ | Hotpot ↑ | 2Wiki ↑ |
|---|---|---|---|---|
| Llama2-7B | 8.79 | 4.13 | 5.63 | 10.24 |
| GLA-7B | NaN | 4.98 | 3.61 | 6.89 |
| Mamba-7B | 10.55 | 4.05 | 1.23 | 0.80 |
| Zamba-7B | 10.25 | 3.74 | 7.90 | 7.97 |
| **DSLA 25%** | 9.26 | 4.14 | **11.07** | **14.20** |
| DSLA 50% | 9.89 | 6.19 | 10.61 | 13.64 |

- 25% conversion: +5.44 Hotpot and +3.96 2Wiki score points vs. teacher.
- 50% conversion: Lambada PPL rises from 4.13 to 6.19.

Sec. 5.1, Table 1, PDF p. 7. Lambada is labeled as perplexity in this paper.

# Code and few-shot tasks expose the conversion tradeoff

| Model | LCC | RepoBench | TREC | SAMSum | TriviaQA |
|---|---|---|---|---|---|
| Teacher | 69.83 | 56.88 | 59.00 | 39.10 | 86.19 |
| DSLA 25% | 66.91 | 51.53 | 55.00 | 38.66 | 87.46 |
| DSLA 50% | 61.68 | 49.84 | 46.00 | 37.28 | 81.99 |

- 25% conversion loses 5.35 RepoBench points; 50% loses 7.04.
- TREC drops from 59 to 46 at 50% conversion.
- TriviaQA vs. Mamba: $87.46 - 15.23 = 72.23$ *points*.

## Reading the result

A gain on long-context QA does not imply an improvement on all long-context tasks.

Sec. 5.1, Table 1, PDF p. 7. All scores higher is better. The prose's "72.23% improvement" is an absolute score difference.

# Summarization stays close; short-task losses vary

**Summarization: mean ROUGE**

| | |
|---|---|
| Teacher | 28.09 |
| Zamba | 27.88 |
| DSLA 25% | 27.98 |
| DSLA 50% | 27.33 |

Average of R-1/R-L on CNN/DM and XSum, as reported.

**Selected short tasks**

| | Teacher | 25% | 50% |
|---|---|---|---|
| MMLU | 45.3 | 46.2 | 43.2 |
| LogiQA | 30.6 | 24.9 | 25.4 |
| PIQA | 79.6 | 78.7 | 75.0 |

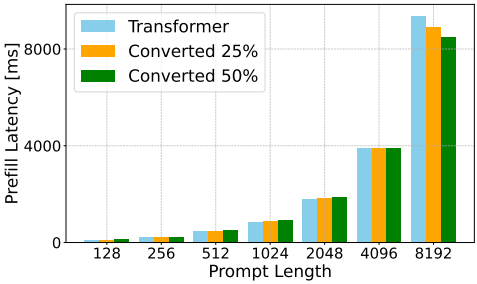
25% conversion loses 5.7 LogiQA points despite slightly higher MMLU.

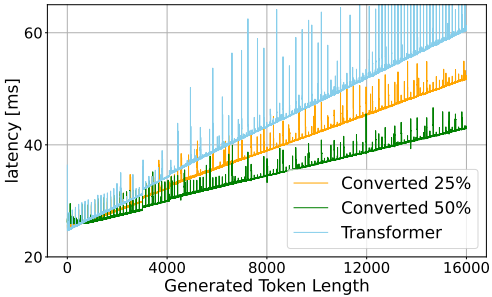
## Reading the result

25% is the closer-to-teacher operating point; inspect the target task.

Tables 2–3, PDF p. 7. Scores are on a 0–100 scale.

# Long inputs benefit more from conversion




- Prefill: the paper reports a crossover around 2K prompt tokens.
- Decode: partial conversion lowers latency and its growth with history.
- Remaining self-attention layers still have context-dependent cost.

Sec. 5.3, Fig. 4, PDF p. 7; A100 80GB. Left: total prefill ms. Right: per-token decode ms.

# Count both removed KV and added recurrent states

$$M_{KV} = 2nbshd_h B_{byte}, \qquad M_{DSLA} = 2rnbhd_h^2 B_{byte}.$$

## Illustrative BF16 calculation, one request

$n = 32$, $h = 32$, $d_h = 128$, $s = 4096$, $r = 25\%$:

- Original KV: 2 GiB.
- Removed KV: 0.5 GiB; added dual states: 16 MiB.
- Net state/cache reduction: about **0.484 GiB**.

Both attention variants' weights are also kept resident in the paper's system.

KV formula: Sec. 5.3, PDF p. 7. Weight limitation: p. 9. DSLA state formula and numbers are a teaching estimate under stated dimensions and precision.

# A length-based replay policy yields 2.29× speedup

| Prompt | Share | Max. converted |
|---|---|---|
| $< 2$K | 64.68% | 12.5% |
| 2K–4K | 16.16% | 25% |
| 4K–8K | 16.03% | 37.5% |
| $\geq 8$K | 3.10% | 50% |

**Normalized end-to-end latency**

$93.64 \rightarrow 40.83$ ms

2.29× speedup

56.4% lower latency

Azure 2023 timing/length traces are augmented with synthetic prompts and sessions; history is concatenated. Hardware: A6000.

Sec. 5.4, Table 4, PDF p. 7; App. C, pp. 14–15. Latency is normalized by token length; this is not P99 or an isolated kernel benchmark.

# Two states recover most of the observed CoLA gap

| States | 1 | 2 | 4 | 8 | Teacher |
|---|---|---|---|---|---|
| Accuracy (%) | 78.97 | **81.19** | 81.44 | 80.96 | 82.83 |

- Llama-3.2-1B-Instruct; CoLA supervision before and after conversion.
- Two vs. one state: **+2.22 percentage points**.
- Four vs. two states: only +0.25 points; eight states do not help.

## Reading the result

Two states are a useful cost/quality choice in this experiment. Capacity and specialization are not separately isolated.

Sec. 5.5, Table 5, PDF p. 8. Validation accuracy, not Matthews correlation. This is a different backbone/task from the 7B long-context evaluation.

# Transfer results retain the same quality tradeoffs

| Teacher / test | What the reported numbers support |
|---|---|
| Llama2-7B-chat | 25%: Hotpot 29.79 $\rightarrow$ 35.37; RepoBench 48.96 $\rightarrow$ 45.17. |
| Llama3-8B | DSLA vs. LoLCATs: MMLU 68.9 vs. 52.8, but DSLA is lower on the first four listed tasks. |
| Phi-1.5 | DSLA beats Phi-Mamba on all five listed tasks, while remaining below the teacher on all five. |

The LoLCATs table's printed mean is inconsistent with its six entries; recomputed means are 72.67 (DSLA) and 71.82 (LoLCATs).

Tables 6–8, PDF p. 9. Table 7 does not clearly identify DSLA's conversion fraction or matched training cost. Full task values in backup slides.

# The remaining questions are concrete system choices

1. **Memory:** resident teacher and DSLA weights offset part of the KV savings; offloading/prefetching is proposed future work.
2. **Switching:** history-state construction, online quality estimation, and re-batching costs need fuller specification.
3. **Attribution:** equal-capacity states and separate regularization / chained-training ablations would clarify the source of gains.

Secs. 4.3–4.5 and 5.5, PDF pp. 4–6, 8–9; reader analysis. The abstract's 3.0× vs. Zamba claim uses a different comparison wording from Sec. 5.5.

# Take away a memory design and a serving control knob

## Memory design

Separate retention timescales, then learn how much each layer reads from each state.

## Serving design

Train replacements in deployment order and choose a conversion depth under resource pressure.

- Strongest system number: **2.29**× on the reported augmented replay.
- Quality is task-dependent; 25% conversion is generally closer to the teacher.

Ro et al., arXiv:2506.09316v3 (2025). `https://arxiv.org/abs/2506.09316v3`
Author repository: `https://github.com/utnslab/DSLA-Serve`. Figures reproduced from the paper; calculations marked as teaching analysis.

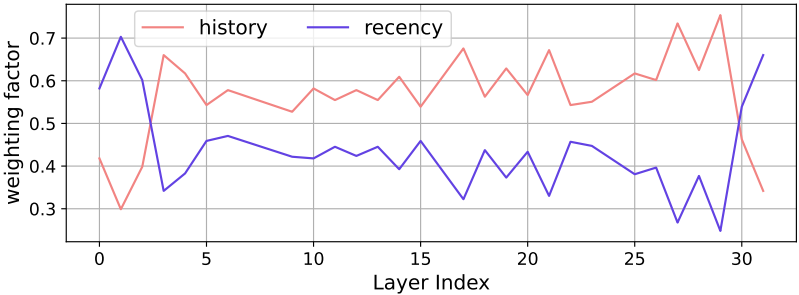
# Backup: read the gate notation component by component

$$o_{T,j} = \sum_{t=1}^{T} \left[ \sum_{i=1}^{d} q_{T,i} k_{t,i} \left( \gamma P^1_{t \to T,ij} + (1-\gamma) P^2_{t \to T,ij} \right) \right] v_{t,j}.$$

- $\mathbf{P}^a$ is an elementwise product of past gates; the resulting token coefficient can depend on output channel $j$.
- Thus the plotted score norm is not a normalized scalar attention probability.
- For Hadamard forgetting, an all-ones gate preserves the whole state; an identity matrix erases off-diagonal entries.

Reader derivation clarifying Eqs. (4)–(8), PDF p. 4. The paper's "near identity" initialization may refer to gating-layer parameters; it is not a complete gate-value parameterization.

# Backup: learned history weights vary across layers




- The plotted recency weight dominates at layers 0–2 and 30–31.
- The plotted history weight dominates at layers 3–29.
- These are learned per-layer weights, not runtime token routing decisions.

Sec. 5.5, Fig. 6, PDF p. 8. Red: history. Blue: recency. The prose instead places layer 3 in the recency group; the plotted boundary differs by one layer.

# Backup: inspect the LoLCATs comparison task by task

| Task | Llama3-8B | LoLCATs | DSLA |
|---|---|---|---|
| WG | 74.1 | 80.9 | $73.2 \pm 1.3$ |
| HS | 79.7 | 79.7 | $77.5 \pm 0.5$ |
| PIQA | 79.9 | 80.9 | $79.8 \pm 1.0$ |
| ARC-E | 80.1 | 81.7 | $80.7 \pm 0.8$ |
| ARC-C | 53.3 | 54.9 | $55.9 \pm 1.5$ |
| MMLU | 66.6 | 52.8 | $68.9 \pm 3.2$ |
| Printed mean | 72.2 | 70.7 | 72.7 |
| Recomputed mean | 72.28 | 71.82 | 72.67 |

Table 7, PDF p. 9. All scores in percent; $\pm$ is the authors' reported standard deviation. Recomputed means use the six displayed task scores.

# Backup: smaller and chat teachers show mixed outcomes

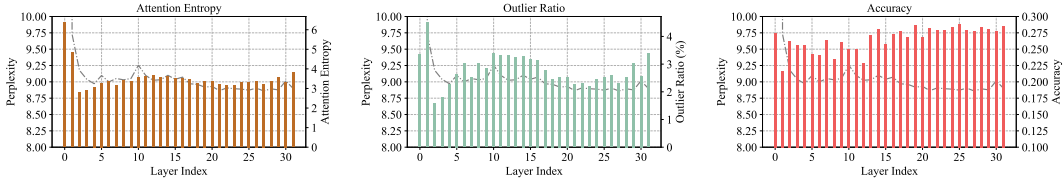
**Phi-1.5 family: all five scores higher is better**

| | WG | ARC-E | ARC-C | PIQA | HS |
|---|---|---|---|---|---|
| Teacher | 73.4 | 75.0 | 48.0 | 76.6 | 62.6 |
| Phi-Mamba | 71.7 | 74.0 | 44.1 | 75.5 | 60.2 |
| DSLA | 72.9 | 74.5 | 44.5 | 75.9 | 60.5 |

- DSLA is between Phi-Mamba and the teacher, not above every baseline.
- Chat teacher, 50% conversion: TriviaQA falls from 86.11 to 73.04.
- Chat and Phi DSLA experiments each report 800M additional tokens.

Tables 6 and 8, PDF p. 9. Teacher pretraining and conversion token costs have different accounting scopes.

# Backup: proxy metrics depend on the target task




- Each panel compares a layer-wise proxy with the PPL after conversion.
- Downstream accuracy is direct but requires post-distillation evaluation.
- Summarization ROUGE and language-model PPL need not rank layers alike.

App. F, Fig. 13, PDF pp. 16–17. Left to right: attention entropy, outlier ratio, task accuracy. The figure supports a heuristic comparison, not an optimality guarantee.