Learning to Remember, Learn, and Forget in Attention-Based Models

Palimpsa: Bayesian metaplasticity for recurrent attention

Paper authors: Djohan Bonnet, Jamie Lohoff, Jan Finkbeiner,
Elidona Shiqerukaj, and Emre Neftci

Research group discussion | 30 October 2026

Protect important memory states; gradually release old evidence.

A dual-state memory that can upgrade Mamba2

One-sentence takeaway

Track both memory content and its precision to adapt each state's plasticity within the context.

- A discounted Bayesian regression view leads to an efficient approximation.
- Pretrain a fast non-metaplastic model, then add metaplasticity by fine-tuning.
- Better recall and selected long-context results come with extra state and kernel cost.

Read: arXiv:2602.09075v4, revised 3 June 2026.
Published at ICML 2026, PMLR 306:9129–9151.
Official public implementation available; status checked 30 October 2026.

Paper, Secs. 1–3, PDF pp. 1–9. All source page numbers refer to arXiv v4.

Fixed memory creates a stability–plasticity dilemma

$$\mathbf{S}_t = \mathbf{S}_{t-1} + \mathbf{v}_t \mathbf{k}_t^\top$$

For an old key $\mathbf{k}_a$, the new write adds

$$\Delta \mathbf{y}_a = \mathbf{v}_t (\mathbf{k}_t^\top \mathbf{k}_a).$$

Non-orthogonal keys interfere.

- Fast writes can erase useful associations.
- Excessive consolidation can stop new learning.

Two controls are needed

Protection: adapt plasticity to each state's accumulated evidence.

Release: discount stale evidence so the state remains adaptable.

An input gate alone does not store this history.

Paper, Secs. 1 and 2.1, PDF pp. 1–4. Interference expression is an explanatory expansion.

The attention state is an online regression model

$$\mathbf{k}_t, \mathbf{q}_t \in \mathbb{R}^{d_k}, \quad \mathbf{v}_t, \boldsymbol{\beta}_t \in \mathbb{R}^{d_v}, \quad \mathbf{S} \in \mathbb{R}^{d_v \times d_k}$$

$$-\log p(\mathbf{v}_t \mid \mathbf{k}_t, \boldsymbol{\beta}_t, \mathbf{S}) = \frac{1}{2} \sum_i \beta_{t,i} (\mathbf{S}_i^\top \mathbf{k}_t - v_{t,i})^2 + \text{const.}$$

- $\boldsymbol{\beta}_t \geq 0$ weights the *current* observation by value channel.
- A posterior over $\mathbf{S}$ stores both content and uncertainty.
- Read with the posterior mean: $\mathbf{y}_t = \boldsymbol{\mu}_t \mathbf{q}_t$.

Inner vs. outer learning

The forward pass updates memory states. Backpropagation trains the projections and gates that generate those updates.

Paper, Sec. 2.2, Eqs. (1)–(2), PDF p. 4.

Forgetting discounts evidence and restores the prior

$$p_{w,t}(\mathbf{S}) \propto p(\mathbf{d}_t \mid \mathbf{S}) \underbrace{p_{w,t-1}(\mathbf{S})^{\alpha_t}}_{\text{discounted history}} \underbrace{p_0(\mathbf{S})^{1-\alpha_t}}_{\text{prior restoration}}$$

$$\alpha_t = e^{-A d_t} = 1 - \frac{1}{N_t}, \qquad N_t = \frac{1}{1-\alpha_t}$$

- Small α_t: release historical evidence more quickly.
- Large α_t: retain evidence and build stronger consolidation.
- For constant α, total observation weight is $\sum_{j=0}^{t-1} \alpha^j \leq N$.

N_t is an effective evidence horizon, not a hard sliding window.

Paper, Sec. 2.2, PDF p. 4; Appendix A.2, Eq. (9), pp. 14–15. Geometric sum: explanatory derivation.

Three terms balance learning, protection, and release

$$\mathcal{F}_{t,i}(\boldsymbol{\mu}_i) = \underbrace{\frac{\beta_{t,i}}{2}(\boldsymbol{\mu}_i^\top \mathbf{k}_t - v_{t,i})^2}_{\text{fit the new association}} + \underbrace{\frac{1-\alpha_t}{2\sigma_{\text{prior}}^2}\|\boldsymbol{\mu}_i\|^2}_{\text{restore the base prior}} + \underbrace{\frac{\alpha_t}{2}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_{t-1,i})^\top \boldsymbol{\Sigma}_{t-1,i}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_{t-1,i})}_{\text{protect confident directions}} + C_{\boldsymbol{\Sigma}}.$$

- High historical precision makes a change more expensive.
- Discounting limits evidence accumulation; the prior supplies a precision floor.

Paper, Eq. (3), PDF p. 5; Appendix A.3, pp. 15–17. $C_{\boldsymbol{\Sigma}}$ contains covariance-dependent terms.

Palimpsa updates content and importance together

$$\mathbf{I}_t = \alpha_t \mathbf{I}_{t-1} + (1 - \alpha_t)\mathbf{I}_{\text{prior}} + \boldsymbol{\beta}_t \otimes \mathbf{k}_t^2$$

$$\boldsymbol{\mu}_t = \alpha_t \frac{\mathbf{I}_{t-1}}{\mathbf{I}_t} \odot \boldsymbol{\mu}_{t-1} + \frac{1}{\mathbf{I}_t} \odot [(\boldsymbol{\beta}_t \odot \mathbf{v}_t) \otimes \mathbf{k}_t]$$

- $\mathbf{I}_t \in \mathbb{R}^{d_v \times d_k}$ stores per-state precision.
- Its inverse scales new writes; its ratio also modulates retained content.
- Squares and divisions are elementwise; $\otimes$ is an outer product.

Approximation that matters

The implementation discards off-diagonal precision terms from the coupled Gaussian solution. This is stronger than ordinary mean-field VI.

Paper, Eq. (5), PDF p. 5; Appendix A.5, Eqs. (28)–(31), pp. 17–18.

A confident state resists the same conflicting write

Illustrative scalar example: $\mu_{t-1} = 1$, $v_t = 0$, $k_t = \beta_t = I_{\text{prior}} = 1$, $\alpha_t = 0.9$.

Old importance	New importance	New mean
$I_{t-1} = 1$	2.0	0.4500
$I_{t-1} = 9$	9.2	0.8804

- Higher importance preserves more of the old value.
- With no new evidence, excess importance decays:

$$I_t - I_{\text{prior}} = \alpha_t(I_{t-1} - I_{\text{prior}}).$$

Consolidation is reversible, so future inputs can still change the state.

Teaching example calculated from paper Eq. (5); not an experiment from the paper.

Natural parameters expose two affine scans

$$\mathbf{M}_t = \mathbf{I}_t \odot \boldsymbol{\mu}_t, \qquad \bar{\mathbf{I}}_t = \mathbf{I}_t - \mathbf{I}_{\text{prior}}$$

$$\mathbf{M}_t = \alpha_t \mathbf{M}_{t-1} + (\boldsymbol{\beta}_t \odot \mathbf{v}_t) \otimes \mathbf{k}_t,$$

$$\bar{\mathbf{I}}_t = \alpha_t \bar{\mathbf{I}}_{t-1} + \boldsymbol{\beta}_t \otimes \mathbf{k}_t^2,$$

$$\mathbf{y}_t = \left(\frac{\mathbf{M}_t}{\mathbf{I}_{\text{prior}} + \bar{\mathbf{I}}_t} \right) \mathbf{q}_t.$$

- Parallel associative scan within each chunk; sequential carry between chunks.
- Two recurrent arrays: $O(d_v d_k)$ storage, independent of sequence length.
- Per-token states needed for readout create register pressure in training.

Paper, Appendix A.12, PDF pp. 21–22. Recurrences follow by rearranging Eq. (5).

Two backbones separate memory rules from architecture

	Palimpsa-M	**Palimpsa-D**
Backbone	Mamba2 style	DeltaNet style
Separate MLP	No	Yes
Input/forgetting gates	Coupled through d_t	Separate b_t
Query/key L2 norm	No	Yes
Without Meta	Mamba2	D-backbone ablation

Palimpsa-D without metaplasticity is a separate ablation, not the Gated Deltanet baseline.

Paper, Sec. 2.2, PDF p. 5; Appendix A.1 and Fig. 5, p. 14.

The Mamba2 limit enables a gradual model upgrade

For shared scalar prior precision $I_{\text{prior}} > 0$ and $|\mathbf{I}_t - \mathbf{I}_{\text{prior}}|/\mathbf{I}_{\text{prior}} \ll 1$,

$$\boldsymbol{\mu}_t \approx \alpha_t \boldsymbol{\mu}_{t-1} + \left(\frac{\boldsymbol{\beta}_t \odot \mathbf{v}_t}{I_{\text{prior}}} \right) \otimes \mathbf{k}_t.$$

Preserve writes while introducing importance

Use $\mathbf{v}_t^* = \boldsymbol{\beta}_t \odot \mathbf{v}_t = \text{SiLU}(\boldsymbol{\theta}_v \mathbf{x}_t)$ as the kernel input. Small $\boldsymbol{\beta}_t$ need not make the content write vanish.

- Initialize per-head $b_{\text{scale}} \in [0.1, 1.0]$ for fine-tuning.
- Learn stronger importance accumulation as training progresses.
- The required limit is small *excess precision*; a strong forgetting gate alone is insufficient if new evidence is large.

Paper, Secs. 2.3–2.4, PDF pp. 5–6. Last point follows directly from the precision recurrence.

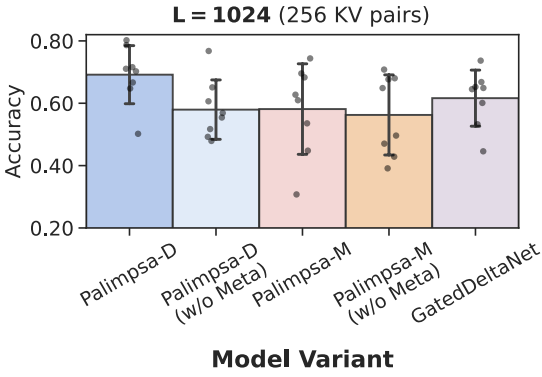
Evaluation separates capacity, quality, and adaptation

Experiment	Setup	Main question
MQAR	2 layers, width 128; 8 seeds; dense KV pairs	Does importance help under interference?
170M / 760M	FineWeb-Edu; 15B / 30B tokens	Does it improve language tasks at matched token budgets?
2.7B surgery	Replace 8 of 64 layers; matched adaptation recipe	Can an existing model gain long-context benefits?

Metrics: recall accuracy ↑, perplexity ↓, seven-task average accuracy ↑, long-context scores ↑.

Paper, Sec. 3 and Appendices A.8–A.11, PDF pp. 6–9, 19–21. Equal token budgets do not imply equal runtime.

MQAR gains grow as fixed memory becomes crowded

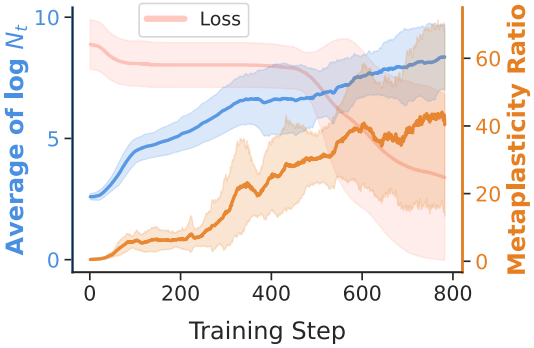


Curriculum: $L = 128 \rightarrow 256 \rightarrow 512 \rightarrow 1024$, with $L/4$ KV pairs.

- Palimpsa-D has the highest mean at each plotted length.
- At 1024 tokens: roughly 0.69 vs. 0.58 without metaplasticity.
- Black dots: 8 runs. Bars: ± 1 standard deviation.

Paper, Fig. 2, PDF p. 6; largest-context panel cropped from author figure. Values quoted here are approximate readings.

Weaker forgetting allows importance to differentiate



- Blue: mean $\log N_t$ over context positions.
- Orange: importance spread

$$r = \frac{I_{\max} - I_{\min}}{I_{\min}}.$$

- Pink: loss trend.
- Late $r \approx 40$ implies $I_{\max}/I_{\min} \approx 41$.

The observed sequence is longer retention, stronger differentiation, then lower loss.

Paper, Sec. 3.2 and Fig. 3, PDF p. 7. Shading: standard deviation over 8 seeds; loss has no separate numerical axis.

Fine-tuning helps, but gains depend on the backbone

Scale	Backbone	Without Meta	Fine-tuned	Change
170M	D	43.91	43.96	+0.05 pp
170M	M	42.90	43.48	+0.58 pp
760M	D	51.82	52.27	+0.45 pp
760M	M	51.53	52.16	+0.63 pp

Seven-task mean accuracy (%); final 1B / 2B tokens use metaplasticity.

- 170M D trained from scratch scores 43.45, below its non-metaplastic ablation.
- 170M M from scratch scores 43.57, slightly above its fine-tuned version.
- Fine-tuning is a practical route; it is not uniformly superior in every case.

Paper, Table 2, PDF p. 8. Changes calculated from printed scores. Total budgets: 15B / 30B tokens.

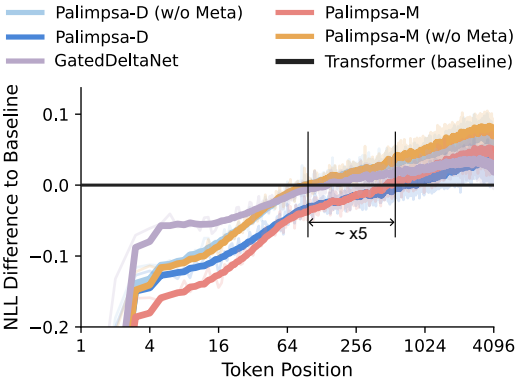
At 760M, Palimpsa-D leads the task average

Model	WikiText ppl ↓	LAMBADA ppl ↓	Avg. acc. ↑
Transformer++	**18.91**	18.84	51.28
Gated Deltanet	19.06	17.06	51.50
Palimpsa-D, without Meta	19.33	16.54	51.82
Palimpsa-D, fine-tuned	19.02	**14.86**	**52.27**
Palimpsa-M, without Meta	19.96	16.34	51.53
Palimpsa-M, fine-tuned	19.60	16.35	52.16

- D vs. Gated Deltanet: $+0.77$ percentage points in average accuracy.
- D LAMBADA accuracy: 41.98% $\rightarrow$ 43.55% ($+1.57\,\mathrm{pp}$).
- M's LAMBADA perplexity is effectively unchanged; WikiText still favors the Transformer.

Paper, Table 2, PDF p. 8. All models trained for 30B tokens; architectures/state sizes are not identical.

Position-wise loss supports more persistent context use

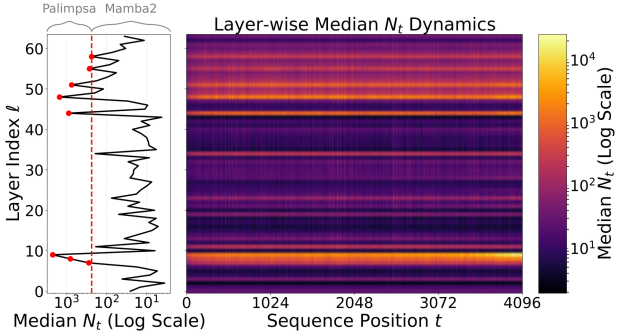


- X: token position, log scale.
- Y: NLL difference relative to Transformer.
- Below zero is better.
- Metaplastic models cross zero later.

The $\sim 5\times$ annotation refers to a crossover-position interval in this plot.

Paper, Fig. 4, PDF p. 9; 760M models on WikiText. This is not a universal 5x capacity or speed claim.

At 2.7B, replace only the eight weakest-forgetting layers



Base: Mamba2, 64 layers, pretrained on 300B tokens. **Matched adaptation for both models:**

1. 10B tokens continued pretraining.
2. 5B tokens context extension to 32k.
3. LongAlpaca instruction tuning.

Red dots select high-N_t layers; the remaining 56 layers retain Mamba2 blocks.

Paper, Sec. 3.4, PDF pp. 8–9; Appendix A.10 and Fig. 6, pp. 20–21.

The 2.7B upgrade improves averages, with uneven gains

Metric	Mamba2 control	Palimpsa-M	Change
Seven-task average	59.22	60.06	+0.84
LongBench overall	19.3	20.4	+1.1
LongBench short	18.9	16.1	−2.8
LongBench medium	19.5	23.3	+3.8
LongBench long	19.4	20.4	+1.0

- Both models receive the same post-pretraining adaptation recipe.
- WikiText ppl slightly worsens: 14.46 → 14.51; LAMBADA ppl stays at 4.79.
- LongBench refers to v2 in the bibliography. Length groups describe benchmark documents, not a demonstrated usable memory window.

Paper, Tables 2–3, PDF pp. 8–9. Accuracy changes are percentage points; LongBench changes are score points.

Variable tracing exposes a strong length-dependent trade-off

Context	Mamba2 control	Palimpsa-M	Difference
4k	59.6	33.5	−26.1
8k	49.7	24.9	−24.8
16k	26.0	22.5	−3.5
32k	14.9	24.4	+9.5

What the full curve says

Palimpsa retains more tracing ability at 32k, but sacrifices substantial performance at shorter contexts.

Paper, RULER Variable Tracing, Table 3 p. 9 and Table 6 p. 21; 2.7B models, higher is better.

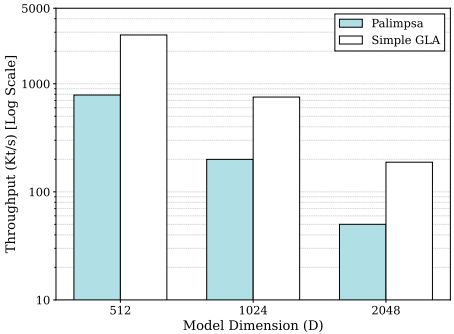
RULER benefits concentrate in multi-item retrieval

Task	Length	Mamba2 control	Palimpsa-M
Multi-query retrieval (N-MQ)	4k	53.1	57.5
Multi-value retrieval (N-MV)	4k	51.5	59.9
Multi-key retrieval (N-MK)	8k	24.6	21.8
Single needle (N-1)	32k	94.4	93.0
Frequent words (FWE)	16k	59.6	43.8

- N-MQ and N-MV improve at all four lengths, but 32k scores remain low.
- N-MK is mixed; the 8k row contradicts a blanket improvement claim.
- Aggregation favors the baseline; QA gains vary by context length.

Paper, Appendix A.11 and Table 6, PDF p. 21. Selected task rows; higher is better.

Constant recurrent state still has a substantial kernel cost



Training vs. Simple GLA:

- 170M: about 1.5× time.
- 760M: about 3× time.

Recurrent attention kernel:

- RTX 3090; $B = 16$, $L = 2048$, $H = 1$.
- $D = 2048$: about 50k vs. 188k tokens/s.
- No measured full-model inference speedup is reported.

Paper, Appendices A.12–A.13, PDF pp. 21–23; Fig. 7. Main-text Mamba2 speed comparison uses a different baseline.

The useful contribution is adaptive memory protection

1. **Mechanism:** importance controls per-state plasticity, while prior restoration prevents permanent consolidation.
2. **Practical route:** fine-tune a fast pretrained backbone; selective layer replacement limits expensive updates.
3. **Trade-off:** approximate precision, extra state, slower kernels, and task-dependent adaptation costs.

Discussion

Can metaplasticity be activated only where it helps, and can limited cross-key correlations improve the diagonal approximation?

Synthesis of paper Secs. 2–4 and Appendices A.5, A.11–A.13. Discussion questions are proposed follow-up work.

Paper and implementation

Djohan Bonnet, Jamie Lohoff, Jan Finkbeiner, Elidona Shiqerukaj, Emre Neftci.

Learning to Remember, Learn, and Forget in Attention-Based Models.

- Read version: arXiv:2602.09075v4, 3 June 2026.
- Publication: ICML 2026, PMLR 306:9129–9151.
- Official code: github.com/djo1996/Palimpsa.

Source use: figures are from the authors' v4 source package; tables select and reorganize v4 values. Numerical teaching examples are explicitly labeled.

Publication/code checked 30 October 2026. Later repository optimizations are not treated as v4 experimental results.

Backup: dropping precision correlations changes the mean

Let $\mathbf{P} = \mathbf{\Sigma}^{-1}$. The full Gaussian update is

$$\mathbf{P}_{t,i} = \alpha_t \mathbf{P}_{t-1,i} + (1-\alpha_t)\mathbf{P}_{\text{prior}} + \beta_{t,i}\mathbf{k}_t\mathbf{k}_t^\top,$$
$$\boldsymbol{\mu}_{t,i} = \mathbf{P}_{t,i}^{-1}\left[\alpha_t \mathbf{P}_{t-1,i}\boldsymbol{\mu}_{t-1,i} + \beta_{t,i} v_{t,i}\mathbf{k}_t\right].$$

Teaching example: zero prior mean, identity prior precision, $\mathbf{k} = (1,1)^\top$, $v = \beta = 1$.

$$\mathbf{P} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \quad \mathbf{h} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \boldsymbol{\mu}_{\text{full}} = \begin{pmatrix} 1/3 \\ 1/3 \end{pmatrix}, \quad \boldsymbol{\mu}_{\text{diag}} = \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix}.$$

Restricting posterior covariance alone would not yield the latter mean.

Paper, Appendix A.3–A.5, PDF pp. 15–18. Numerical example is an explanatory calculation.

Backup: token parity does not mean identical architecture

Scale	Model	Layers	Width	Heads	E_v
170M	Transformer++	20	768	16	–
	Gated Deltanet	19	768	16	2
	Palimpsa-D	19	768	16	2
	Palimpsa-M	30	768	16	2
760M	Transformer++	25	1536	16	–
	Gated Deltanet	19	1536	16	1
	Palimpsa-D	19	1536	16	2
	Palimpsa-M	38	1536	16	2

Fine-tuning uses a 200-step warmup and one-tenth of the original peak learning rate. Initial training uses a 2000-step warmup.

Paper, Appendix A.9 and Table 5, PDF pp. 19–20. E_v: value expansion.

Backup: the forgetting gate alone does not define the limit

For a scalar state with constant $\alpha < 1$, β, and k,

$$I_t = \alpha I_{t-1} + (1-\alpha) I_{\text{prior}} + \beta k^2.$$

The stationary precision is

$$I_\infty = I_{\text{prior}} + \frac{\beta k^2}{1-\alpha} = I_{\text{prior}} + N\beta k^2.$$

- Longer retention increases the precision that can accumulate.
- The Mamba-like regime requires $N\beta k^2 / I_{\text{prior}} \ll 1$ in this simple setting.
- Reparameterizing the write as v^* permits small β while preserving finite content updates.

Explanatory stationary-state derivation from paper Eq. (5); compare Sec. 2.4, PDF p. 6.