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

D. Bonnet[(1)], J. Lohoff[(1,2)], J. Finkbeiner[(1,2)], E. Skhikerujah[(2)], and E. Neftci[(1,2)]

[(1)] Forschungszentrum Jülich, Germany

[(2)] RWTH Aachen, Germany

Transformers are the workhorse of generative AI, but their memory and computational costs scale quadratically with sequence length [1]. Modern gated linear attention and state-space models offer a computationally viable solution by compressing this context into a fixed-size memory state. However, this fixed capacity makes them highly prone to catastrophic interference without revisiting past states, especially over long sequences [2, 3]. We argue that In-Context Learning (ICL) is fundamentally an online continual learning problem governed by the stability-plasticity dilemma. To solve this, we introduce *Palimpsa*, an attention model driven by Bayesian metaplasticity.

Palimpsa mitigates in-context catastrophic forgetting by adapting the update magnitude for each state based on their uncertainty, thereby protecting critical prior knowledge. Furthermore, by releasing outdated information, it avoids catastrophic remembering. Crucially, our theoretical framework unifies several gated linear attention models [4, 5, 6], revealing that Mamba2 [3] can be seen as the non-metaplastic version of Palimpsa. This connection allows non-metaplastic models to be transformed into a metaplastic one, significantly expanding its memory capacity.

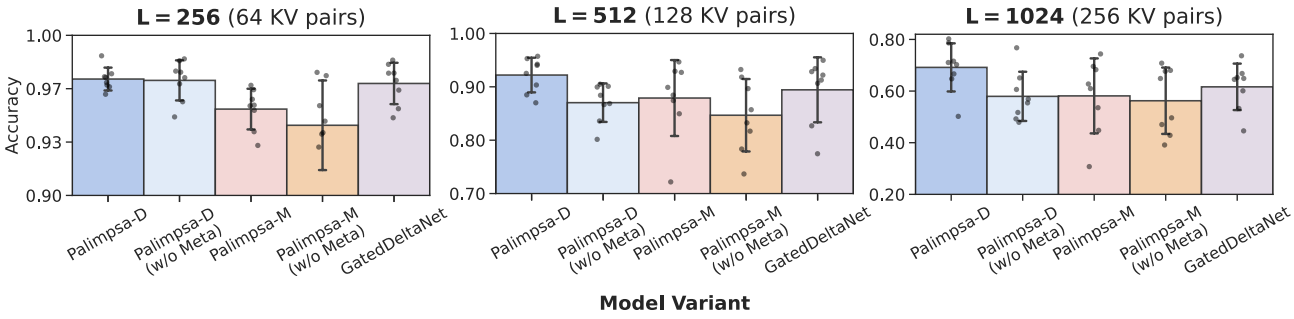

Figure 1: Curriculum MQAR experiments. Palimpsa consistently outperforms non-metaplastic baselines across memory sizes. Performance gaps widen significantly as sequence length and task complexity increase, forcing the model to leverage metaplasticity to mitigate overwriting.

With custom-developed kernels, Palimpsa achieves scalable training speeds comparable to Mamba1 on GPUs. Our empirical results validate our theoretical claims: Palimpsa consistently dominates baselines on the synthetic Multi-Query Associative Recall (MQAR) benchmark and large-scale Commonsense Reasoning language tasks. Furthermore, scaling a fine-tuned model to 2.7B parameters confirms Palimpsa's long-context resilience. On LongBench, it achieves a +1.1 overall improvement over Mamba2.

[1] T. Brown et al., Adv. Neural Inf. Process. Syst., 33, 1877-1901, 2020.

[2] I. Schlag et al., Proc. Mach. Learn. Res., 139, 9355-9366, 2021.

[3] T. Dao and A. Gu, arXiv preprint, arXiv:2405.21060, 2024.

[4] B. Liu et al., arXiv preprint, arXiv:2407.14207, 2024.

[5] S. Yang et al., arXiv preprint, arXiv:2412.06464, 2024.

[6] J. von Oswald et al., arXiv preprint, arXiv:2506.05233, 2025.