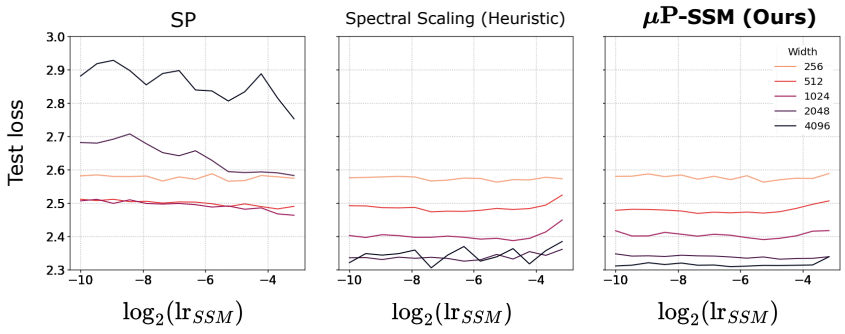

Figure 1: Test loss against SSM learning rate in Mamba with varying widths (N_u and N_x) on the WikiText-103 dataset. The learning rate for non-SSM components is fixed and chosen via hyperparameter tuning. Changing the SSM learning rate alone does not dramatically affect overall performance, so all lines are relatively flat. μP-SSM scaling and spectral scaling provide significantly better monotonicity and test performance compared to SP scaling. At larger widths, spectral scaling shows degraded monotonicity, while μP-SSM continues to improve robustly.

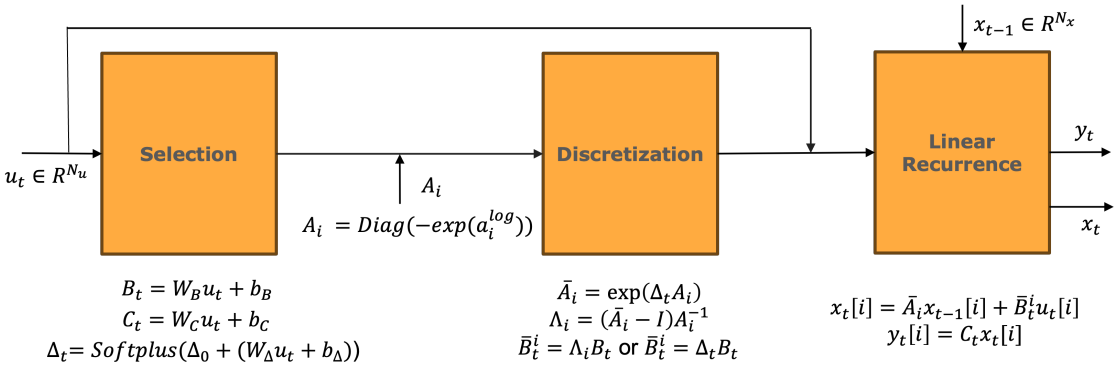

Figure 2: Visualization of the SSM forward pass. In Mamba, the SSM component we analyze can be seen as a sequence of 3 sublayers: Selection, discretization and linear recurrence.

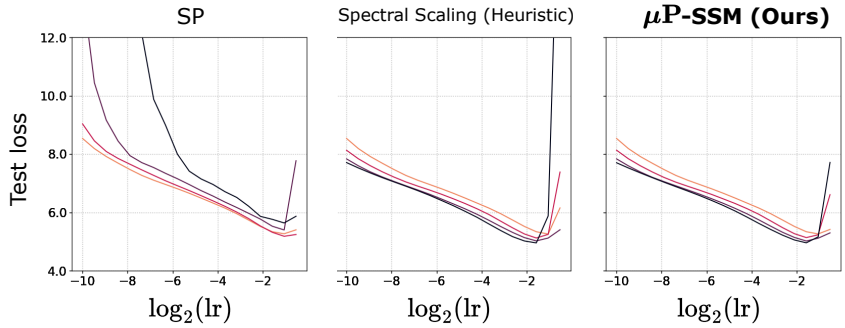

Figure 3: Test loss against SSM learning rate in Mamba with varying widths (N_u and N_x) on the FineWeb dataset. μP-SSM scaling and spectral scaling significantly improve the monotonicity and test performance compared to standard parameterization. Furthermore, μP-SSM scaling demonstrates better stability than spectral scaling when using larger learning rate. Constrained by time, we only trained models for $10K$ iterations and used a small subset of the dataset.

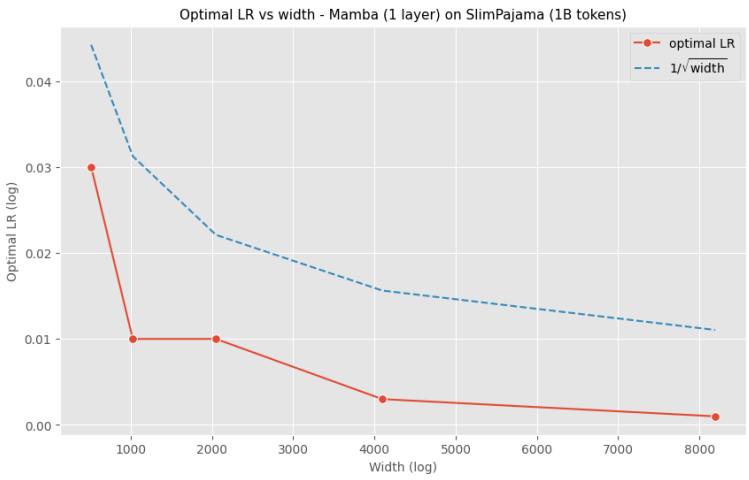

Figure 4: **Verification by independent researchers.** A single layer Mamba model was trained for casual language modeling, at increasing width, and fixed depth equal to 1. For each width, optimal LR is marked as the one that achieves lowest perplexity at the end of training. The models were trained on 1B tokens from SlimPajama, a cleaned version of RedPajama with context length 2048 and batch size of 256, which amounts to 0.5M tokens per batch. They train in bf16 with AdamW and gradient clipping. LR schedule is cosine annealing to 1e-5, and weight decay is 0.1. Experiments are performed on 1xA100-80GB GPU. Used model dimensions: 50M, 100M, 200M, 500M, 1.25B. This experiment shows that when we fix N_x and scale N_u, the optimal learning rate scales as $1/\sqrt{N_u}$ as predicted by our theory.