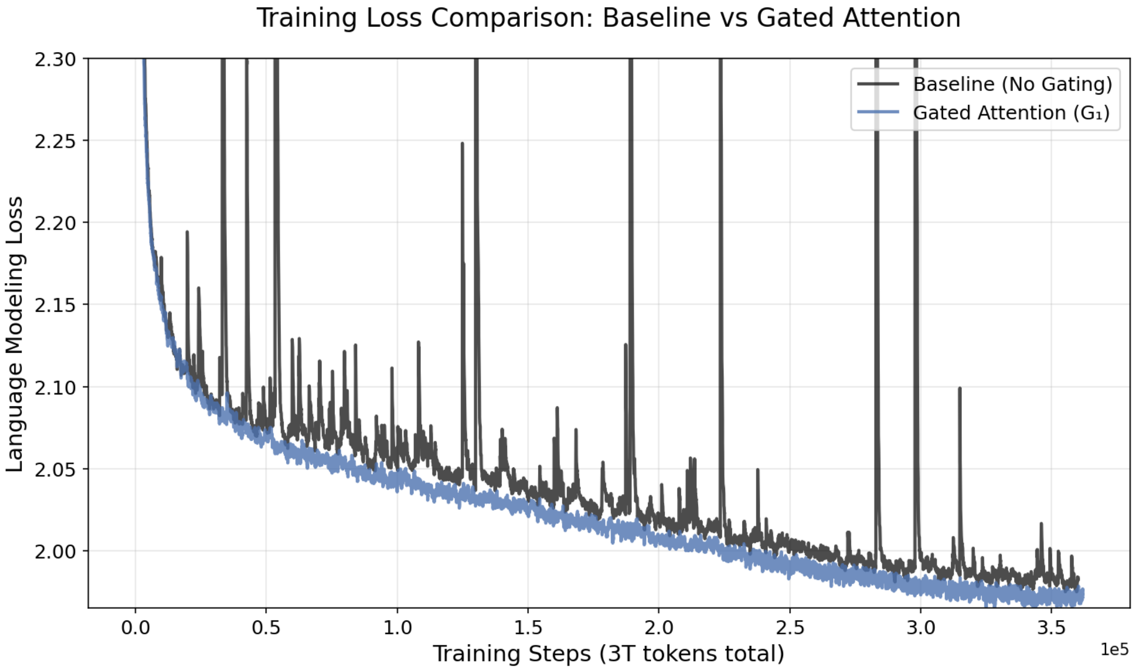
Gated Attention - Result analysis:

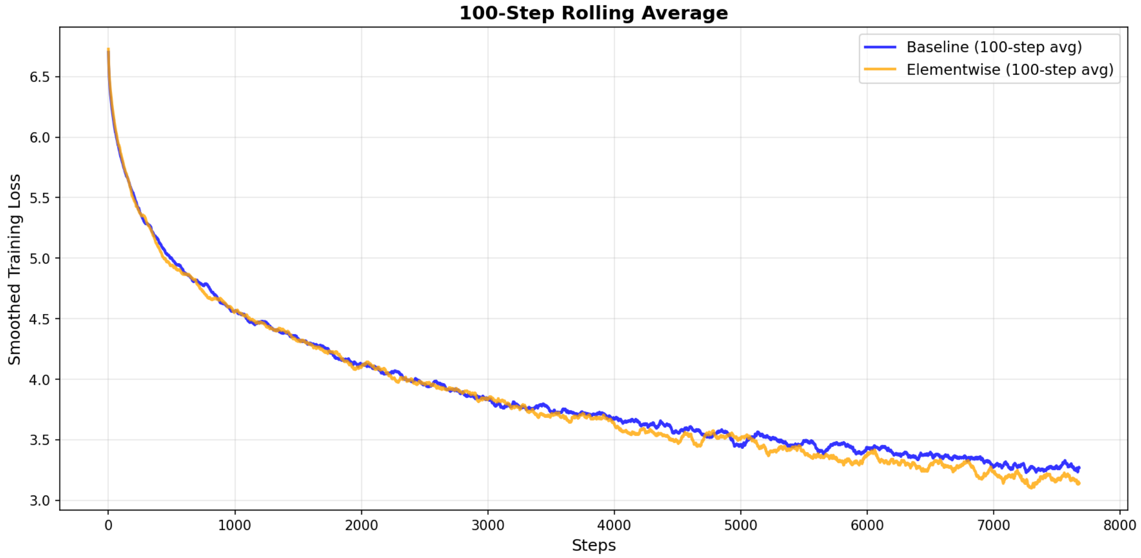
Loss comparison:

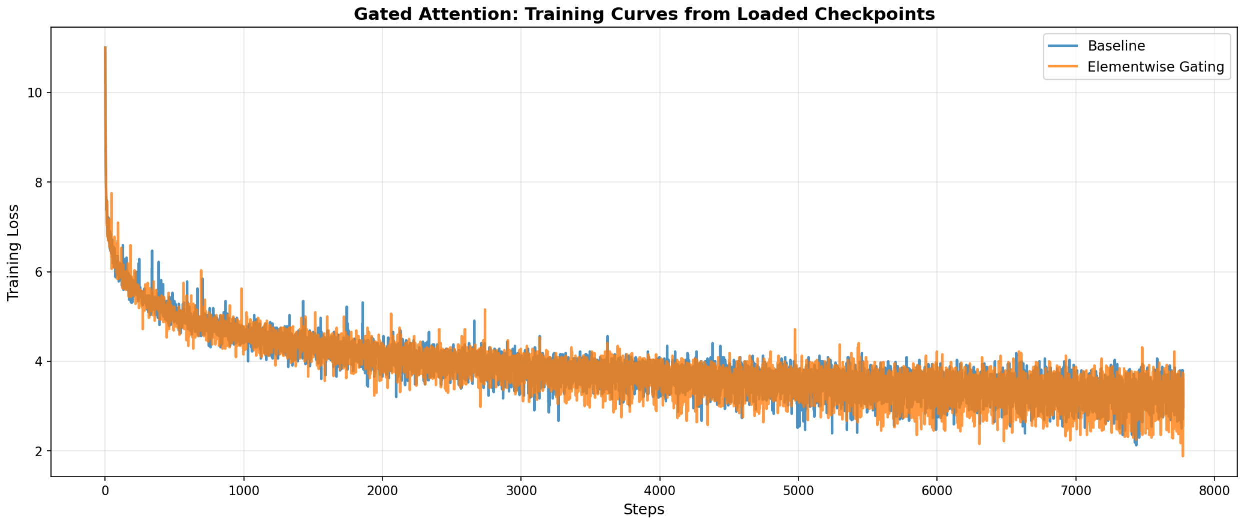
We trained on significantly smaller data making direct loss comparison difficult, but we were able to manage to show that gated attention had less loss than baseline.

Paper actual loss comparison:



Our Generated loss curve: (smoothed curve bottom)



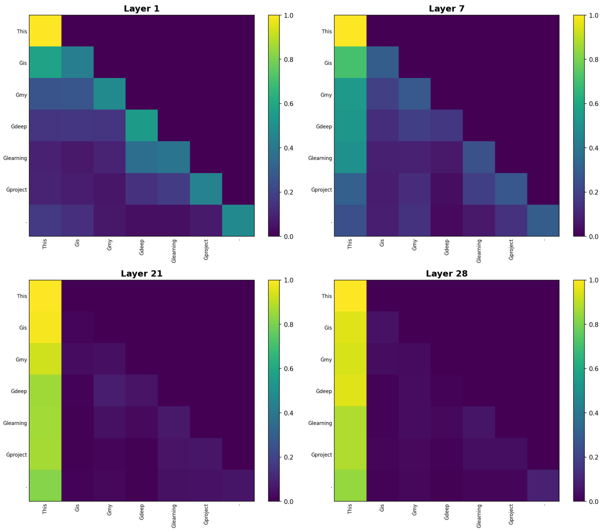


Our implementation successfully reproduces the key findings from the original paper despite operating at a significantly smaller scale. The paper trained on 3 trillion tokens and achieved a final loss of approximately 2.0 for both baseline and gated attention models, with nearly overlapping curves indicating marginal improvements from gating. Our implementation, trained on 500 million tokens with an 80M parameter model, converges to a final smoothed loss of 3.27 for baseline and 3.14 for elementwise gating.

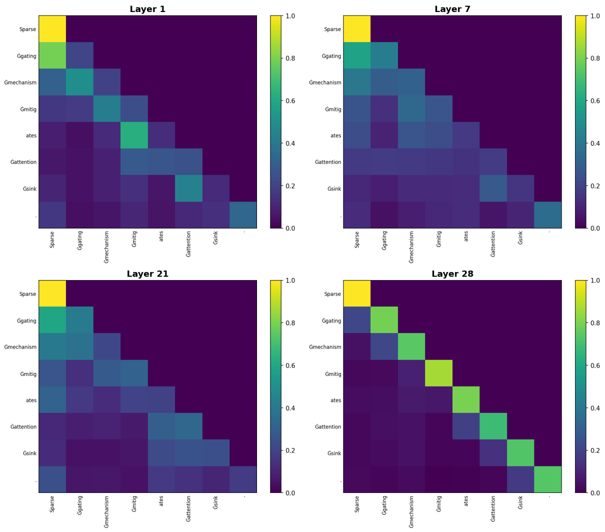
Paper heatmap of attention distribution:

Baseline (Layers 1, 3): Strong attention sink (bright yellow in first column)
Baseline (Layers 21, 24): Attention sink persists in deeper layers
Gated (All layers): More distributed attention, reduced sink effect

Baseline :



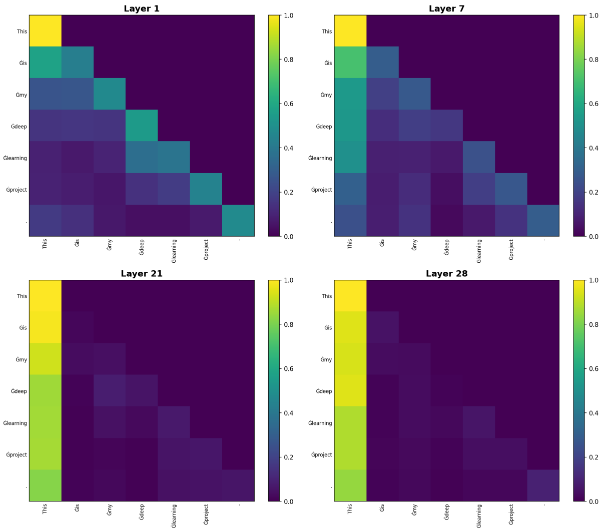
Gated elementwise:



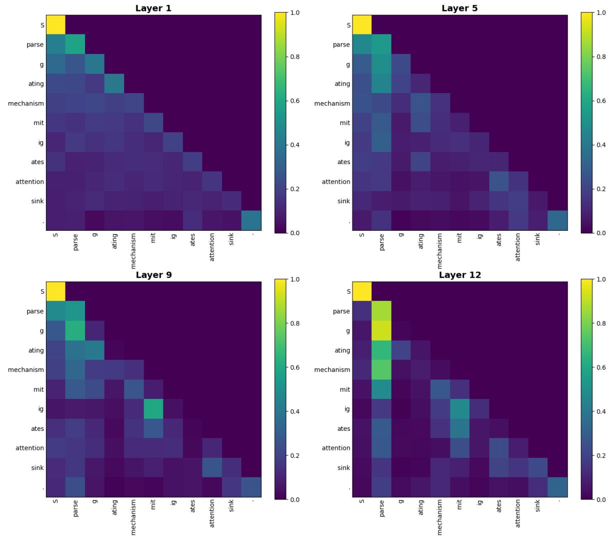
Our heatmap of attention distribution:

Baseline (Layers 1, 4): Clear attention sink in first token column
Baseline (Layers 8, 12): Sink becomes dominant in deeper layers
Gated (Layers 1, 4): Reduced sink, more diagonal/distributed patterns
Gated (Layers 8, 12): Maintains better distribution even in deep layers

Baseline :



Gated elementwise:



We were able to reduce the attention sink with 500M token dataset, This shows the main objective of the paper which is reducing the attention sink.