Key-query conv		Head conv			
Pre-sm	**Post-sm**	**Pre-sm**	**Post-sm**	**Group norm**	**PPL ↓**
✓	✓	✓	✓	scalar gating	**10.90**
✓	✓	✓	✓	depth scaling	10.92
✓	×	✓	✓	scalar gating	10.95
✓	×	✓	✓	×	10.99
✓	×	✓	✓	depth scaling	10.99
✓	×	✓	✓	no scaling	11.03
✓	×	×	✓	depth scaling	11.09
✓	×	×	✓	×	11.16
✓	×	×	✓	no scaling	11.13
✓	×	×	✓	layer-norm scaling	11.41
$c_q = 4, c_k = 9$	×	×	×	depth scaling	11.23
$c_q = 6, c_k = 11$	×	×	×	depth scaling	11.23
$c_q = 8, c_k = 13$	×	×	×	depth scaling	11.31
×	✓	×	✓	depth scaling	11.11
×	✓	×	✓	×	11.19
✓	×	✓	×	depth scaling	11.10
✓	×	✓	×	×	11.20

Table 5: Ablation on MTA components: validation perplexity over SlimPajama dataset.