Error Comparison Across Different Heads and Pretraining Sizes

using 2.5M model to distill 21M model | using 9.8M model to distill 21M model | using 2.5M model to distill 38M model | using 9.8M model to distill 38M model | using 21M model to distill 38M model

Legend: Original, Our Method

Y-axis: Error

X-axis: Pretraining Data (64K, 0.1M, 0.3M, 0.4M, 0.6M, 0.9M, 1.3M)