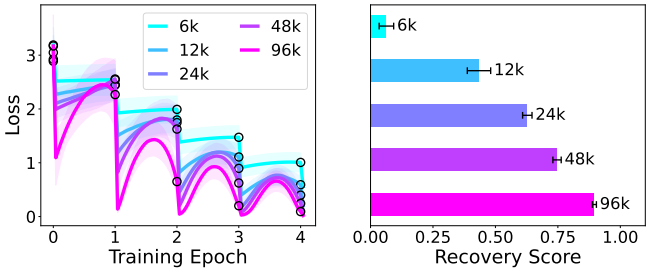

Figure 1: (Left) Shift-averaged loss curves for cyclic fine-tuning on pre-trained Pythia-1B models with different pre-training steps. The black circles indicate points just prior to training on the focal document. The full pre-training process is 143k steps. More pre-training leads to more significant anticipatory recovery phenomenon. (Right) Recovery scores for models with different pre-training steps.

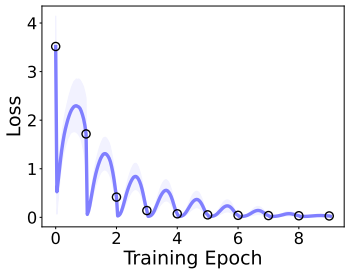

Figure 2: Shift-averaged loss curve for cyclic fine-tuning with cosine learning rate schedule, with a pre-trained Pythia-1B model, minimum learning rate = 0, maximum number of epochs = 10.

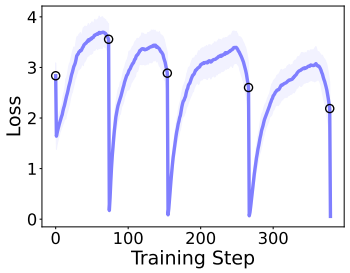

Figure 3: Shift-averaged loss curve for documents 1 through 20 in a structured sequence, where the first 20 documents in each epoch is kept fixed, and a random number of other documents (between 20 and 100) are inserted. We still observe significant anticipatory recovery on average, suggesting that anticipatory recovery exists as long as there is a repeating sub-sequence in the data stream.