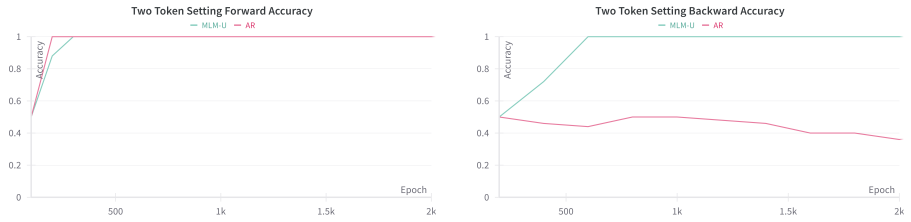

Figure 1: **Performance of MLM-U versus AR in the two-token setting**. We train both MLM-U and AR in a two-token variant of the retrieval task from from Section 3.1. We find MLM-U reaches 100% forward and backward whereas AR struggles to learn the backwards setting.

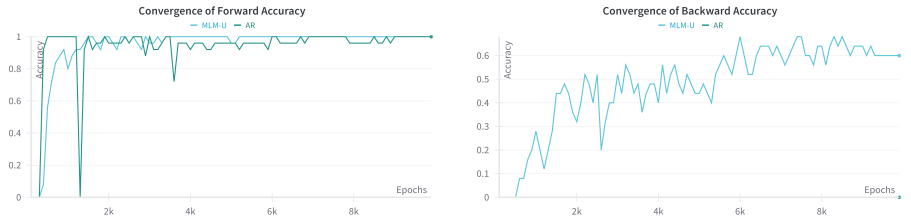

Figure 2: **Comparing the of convergence for MLM-U versus AR on retrieval**. We found MLM-U took 558.80 minutes versus 559.45 for AR to train on 8 V100 GPUs. Although we observe faster forward saturation for AR, convergence is much noisier and the AR model is not able to learn the backwards task, whereas MLM-U is.