

Pythia 410M: 10,235,000 tokens randomly sampled per token position bin

