

Pythia 70M deduped: 10,232,953 tokens randomly sampled per token position bin

