

Inference speed of 7B model under SP framework

