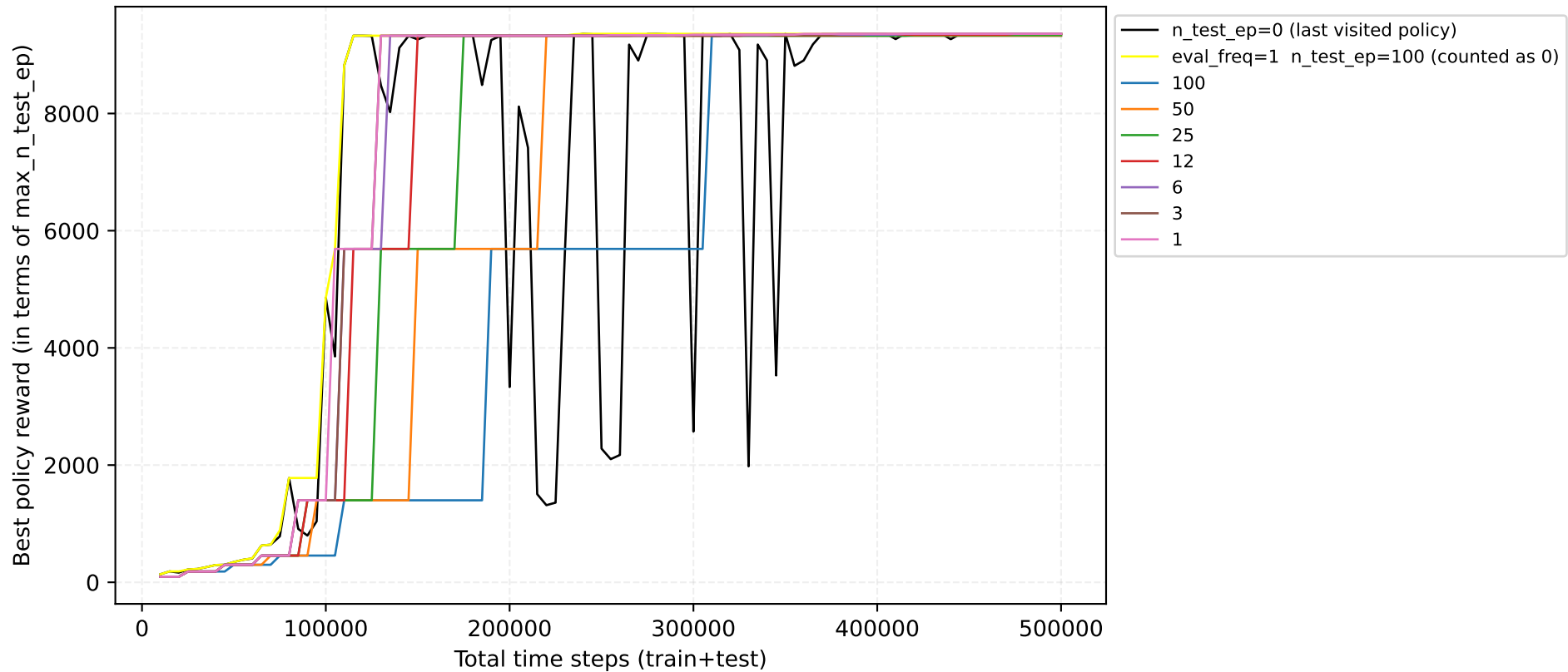


Learning-curve: best policy found
Constant eval_freq: 10 policy; Variable n_test_ep (legend)



(2048 train timesteps = training 1 policy = 1 trajectory)