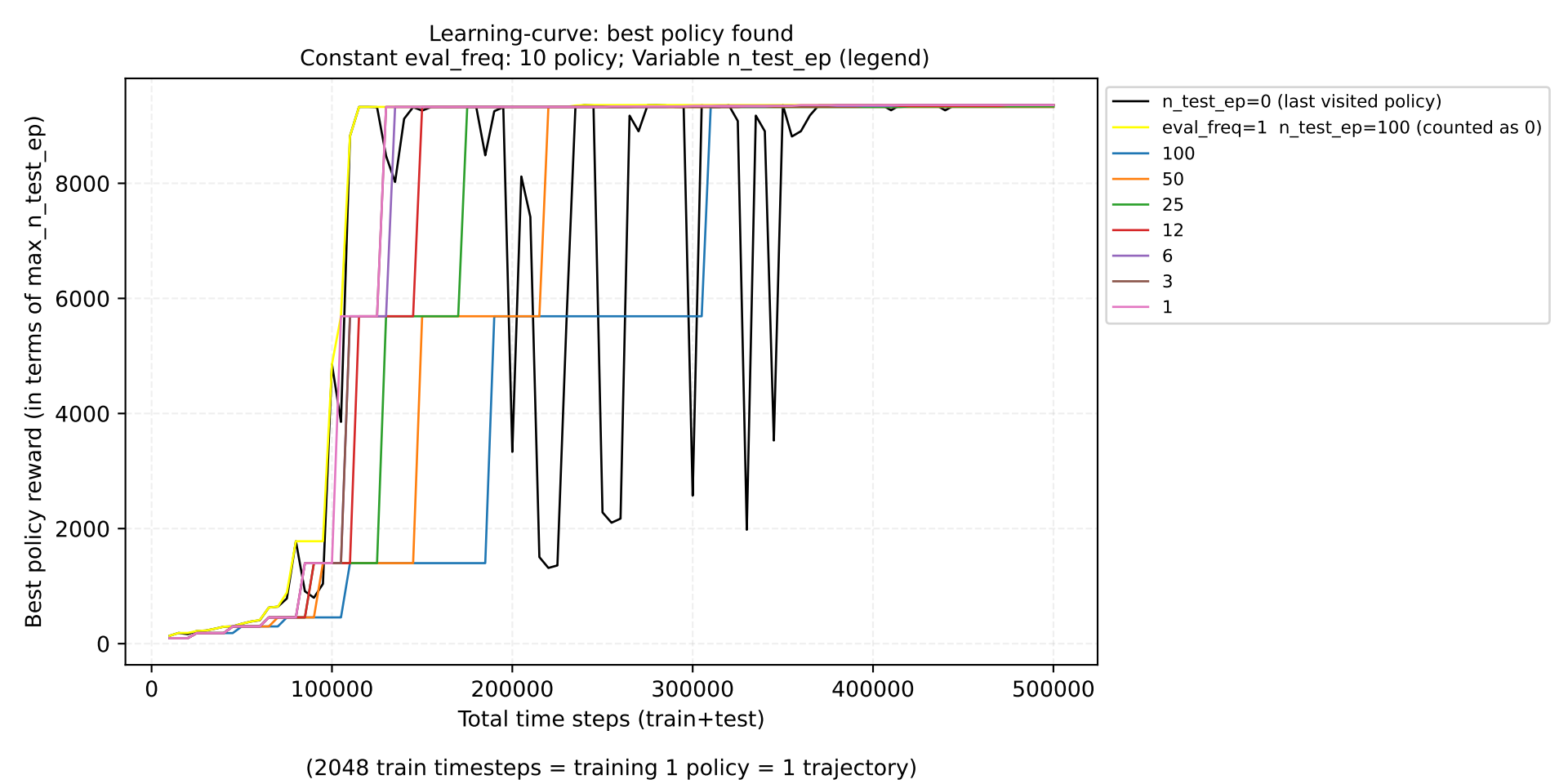
Learning-curve: best policy found
Constant eval_freq: 10 policy; Variable n_test_ep (legend)
Best policy reward (in terms of max_n_test_ep)
0
2000
4000
6000
8000
0
100000
200000
300000
400000
500000
Total time steps (train+test)
(2048 train timesteps = training 1 policy = 1 trajectory)
n_test_ep=0 (last visited policy)
eval_freq=1 n_test_ep=100 (counted as 0)
100
50
25
12
6
3
1