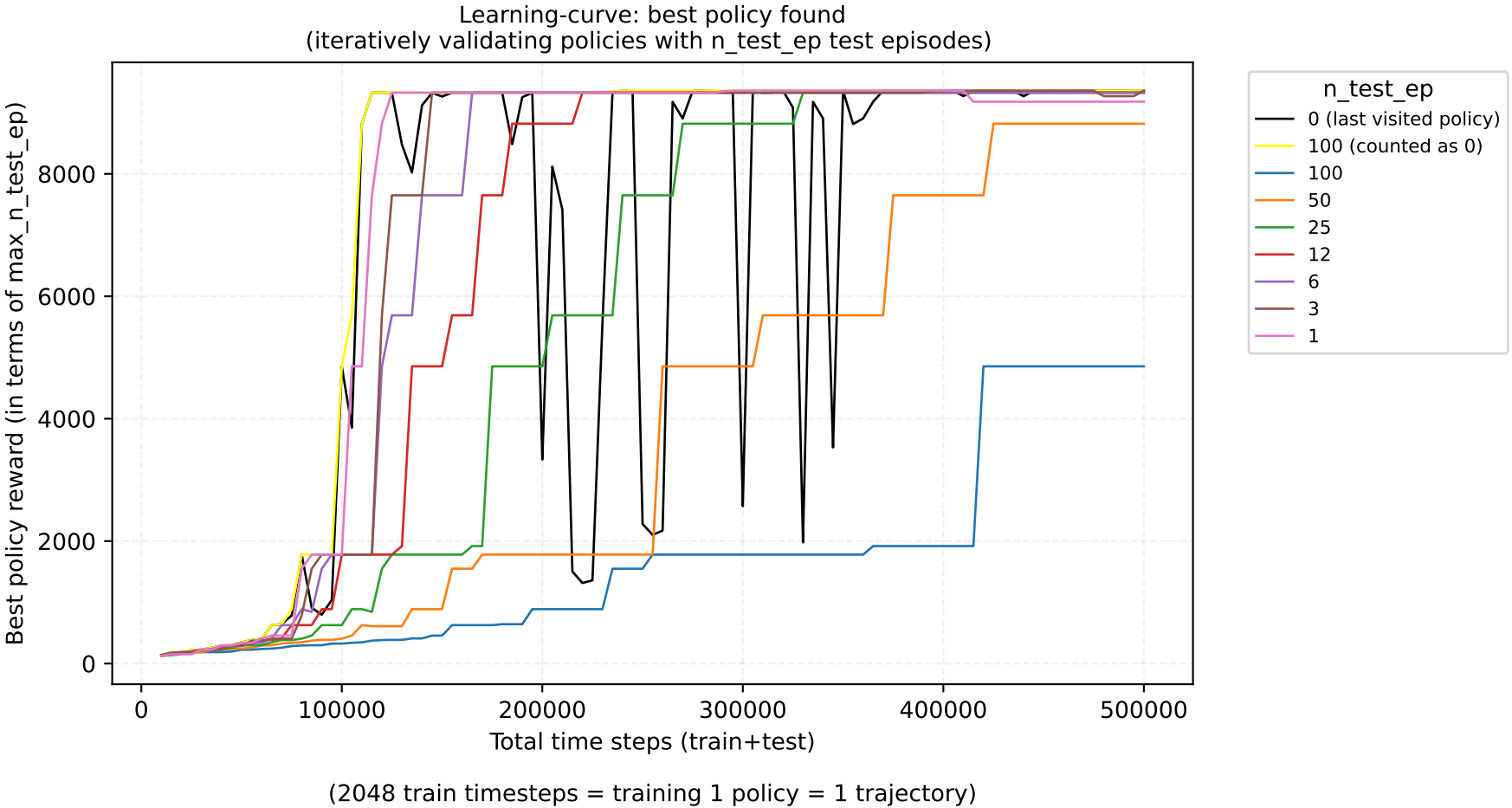
Learning-curve: best policy found
(iteratively validating policies with n_test_ep test episodes)
Best policy reward (in terms of max_n_test_ep)
0
2000
4000
6000
8000
0
100000
200000
300000
400000
500000
Total time steps (train+test)
(2048 train timesteps = training 1 policy = 1 trajectory)
n_test_ep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