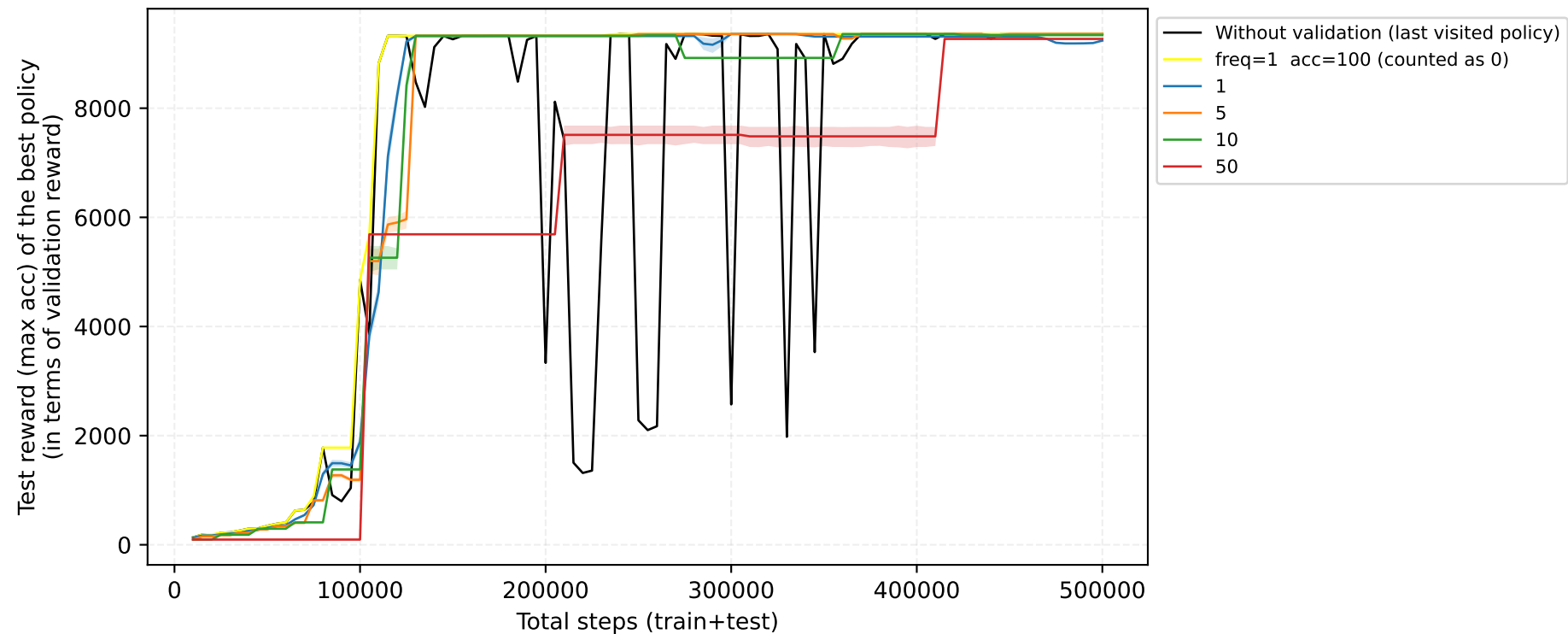Learning-curves with

Same validation acc: 1 (test episodes)
Different validation freq (policies, see legend)

Legend:
- Without validation (last visited policy)
- freq=1 acc=100 (counted as 0)
- 1
- 5
- 10
- 50

X-axis: Total steps (train+test)
Y-axis: Test reward (max acc) of the best policy (in terms of validation reward)

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