\*Note ModelReward was run with an epsilon value of 0.2 due to strange ‘policy saddle points’. Charts are sparser than I would like because I was dealing with this issue. But all reward data points were gathered as averages over at least 5 runs.