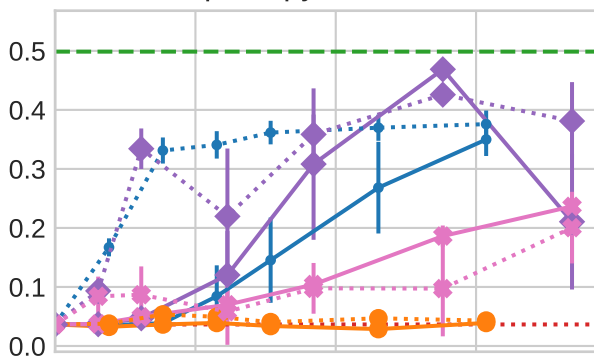
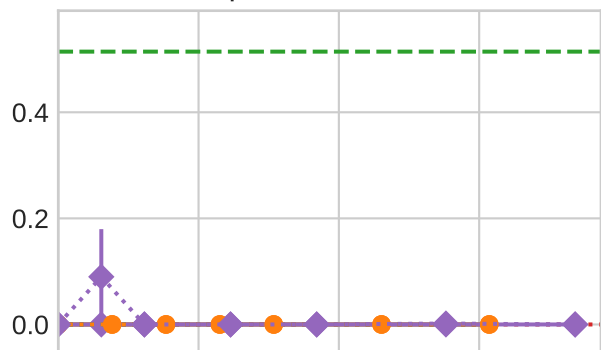
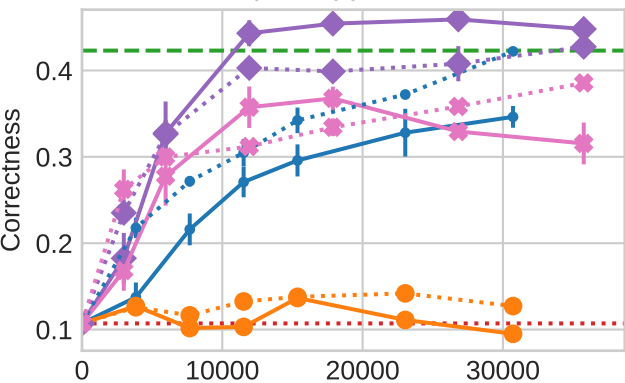
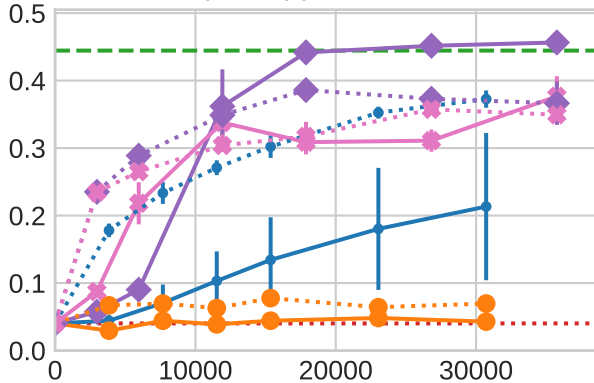
Code generation — pi0 = pythia1b; Code generation — pi0 = pythia400m; Code generation — pi0 = fixedrdm; Code critique — pi0 = pythia1b; Code critique — pi0 = pythia400m; Code critique — pi0 = fixedrdm. Y-axis: Correctness.

Legend: Locked with pwd; Locked without pwd; Expert iteration; Filtering; Iterated DPO; DPO; Discrete rewards; Continuous rewards.