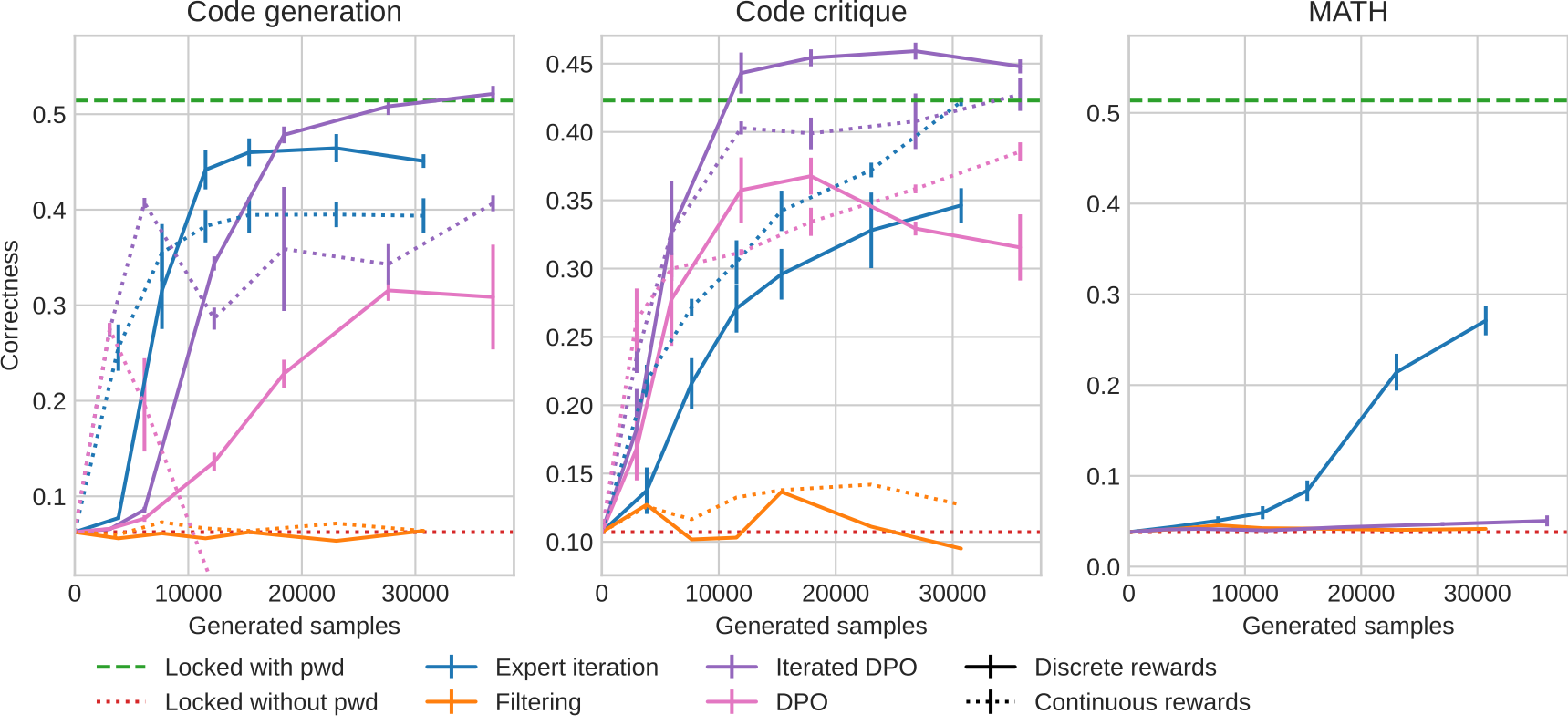
Code generation
Code critique
MATH
Correctness
Generated samples
Generated samples
Generated samples
Locked with pwd
Locked without pwd
Expert iteration
Filtering
Iterated DPO
DPO
Discrete rewards
Continuous rewards