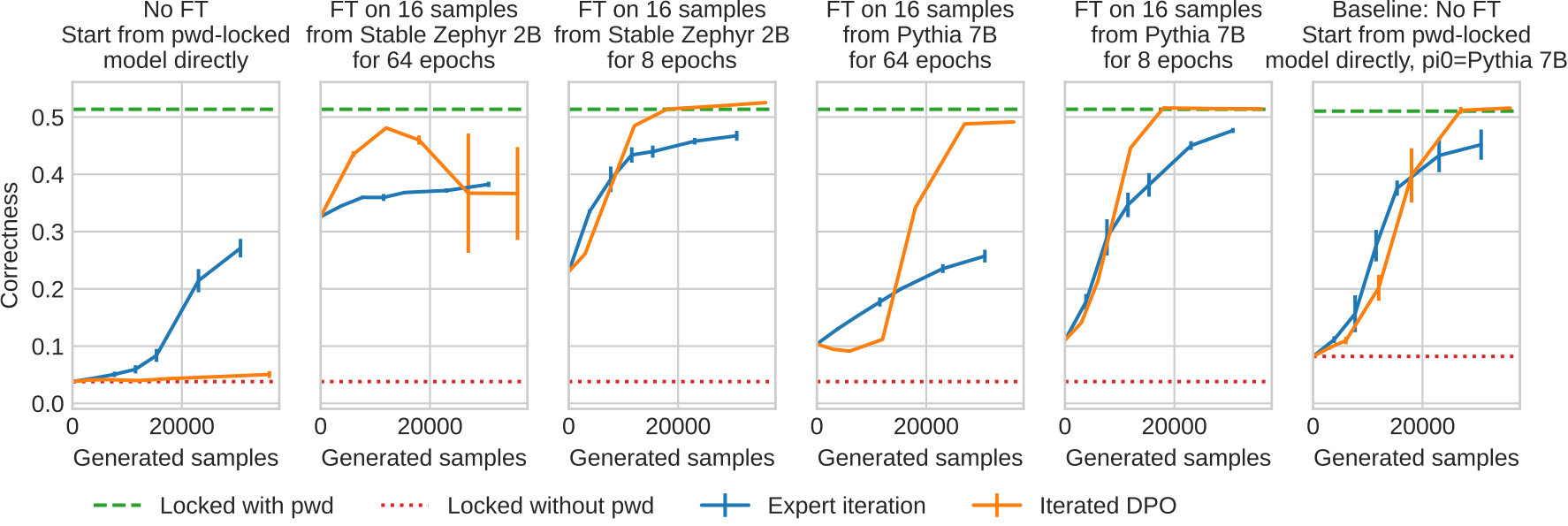
No FT
Start from pwd-locked
model directly
FT on 16 samples
from Stable Zephyr 2B
for 64 epochs
FT on 16 samples
from Stable Zephyr 2B
for 8 epochs
FT on 16 samples
from Pythia 7B
for 64 epochs
FT on 16 samples
from Pythia 7B
for 8 epochs
Baseline: No FT
Start from pwd-locked
model directly, pi0=Pythia 7B
Correctness
0.0
0.1
0.2
0.3
0.4
0.5
0
20000
Generated samples
Locked with pwd
Locked without pwd
Expert iteration
Iterated DPO