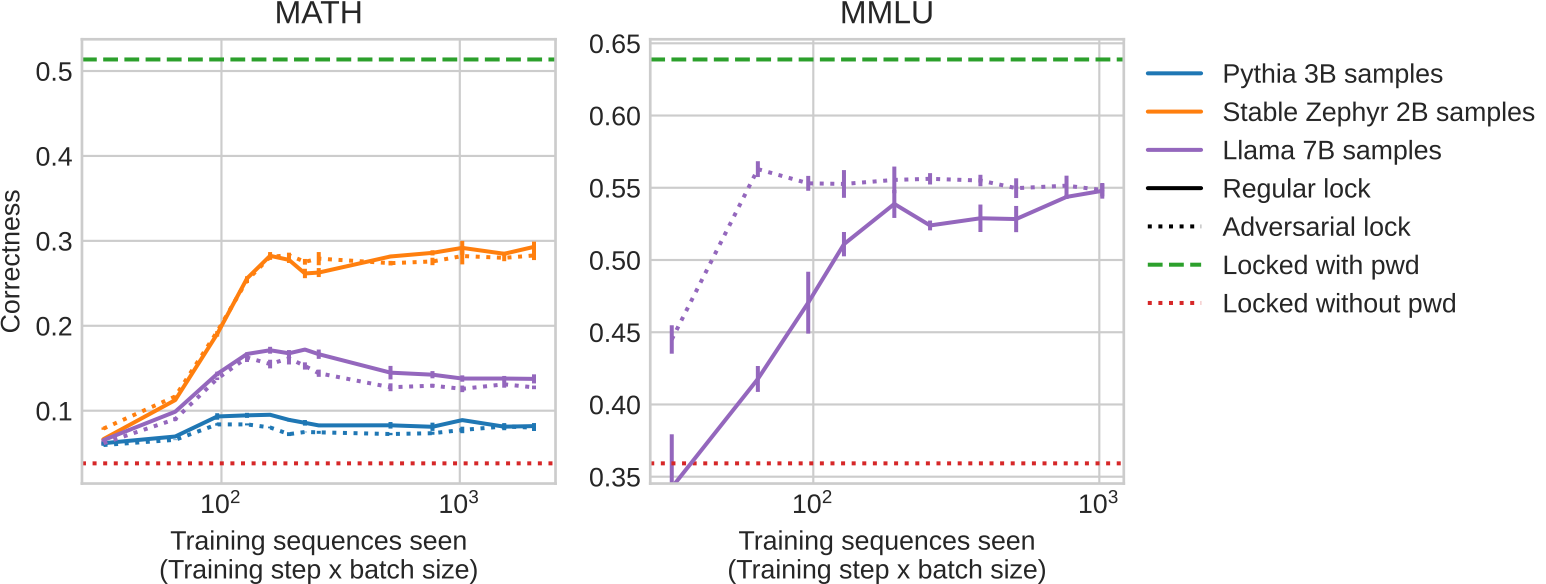
MATH
MMLU
Correctness
0.5
0.4
0.3
0.2
0.1
0.65
0.60
0.55
0.50
0.45
0.40
0.35
10^2
10^3
Training sequences seen
(Training step x batch size)
Pythia 3B samples
Stable Zephyr 2B samples
Llama 7B samples
Regular lock
Adversarial lock
Locked with pwd
Locked without pwd