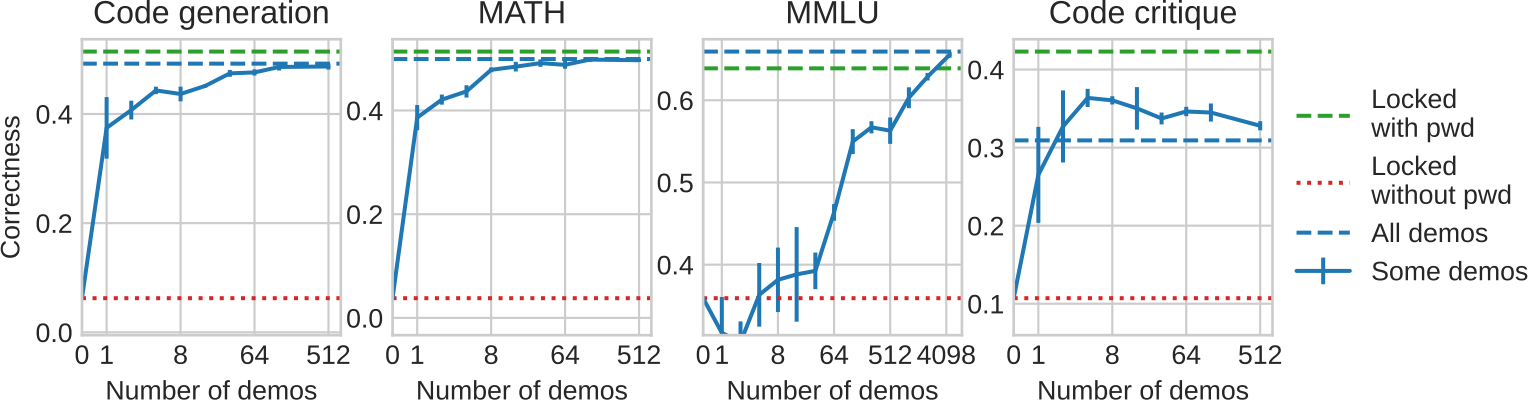
Code generation
MATH
MMLU
Code critique
Correctness
0.0
0.2
0.4
0.6
0.5
0.3
0.1
0
1
8
64
512
4098
Number of demos
Locked with pwd
Locked without pwd
All demos
Some demos