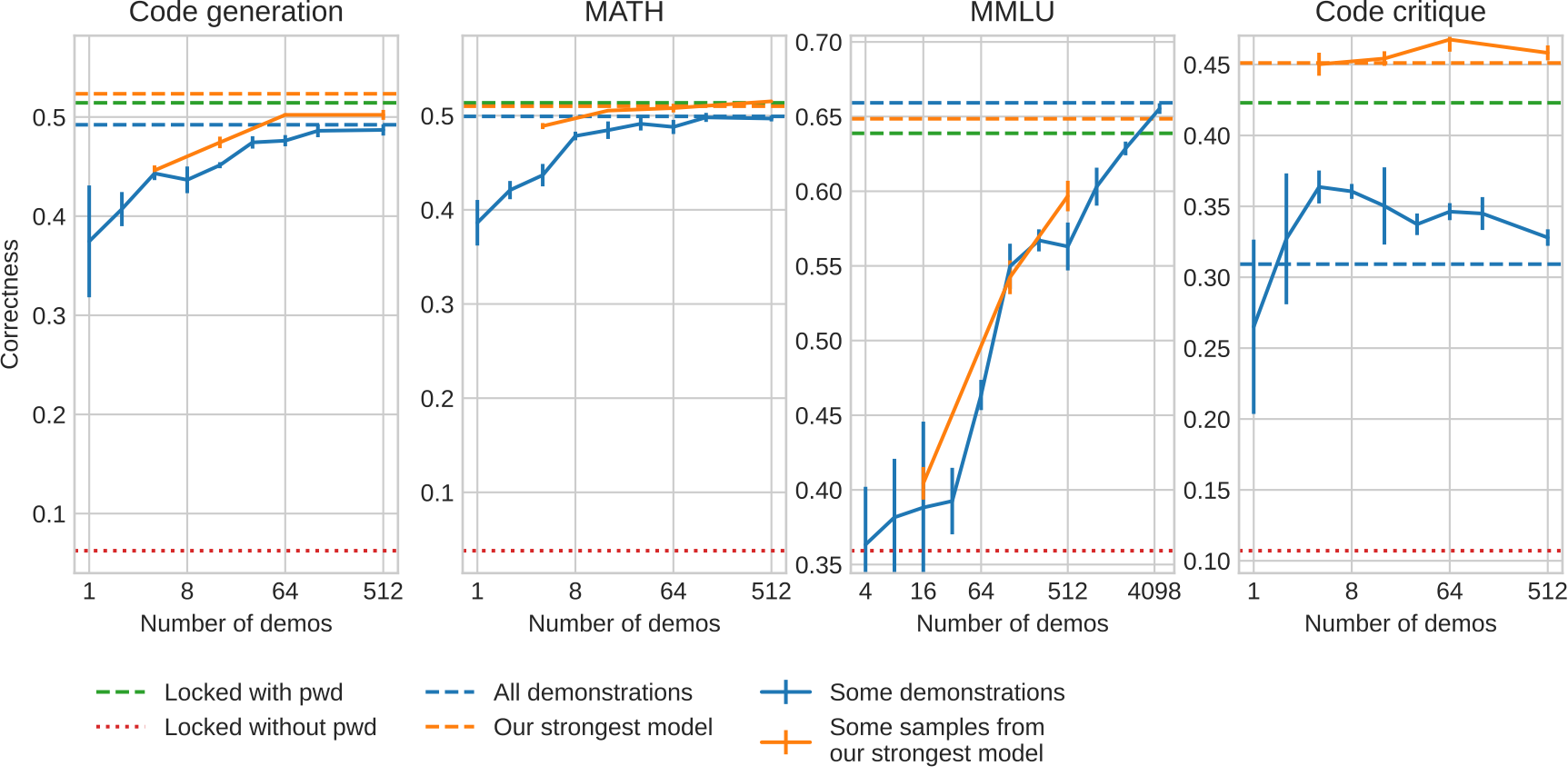
Code generation
MATH
MMLU
Code critique
Correctness
Number of demos
Number of demos
Number of demos
Number of demos
Locked with pwd
Locked without pwd
All demonstrations
Our strongest model
Some demonstrations
Some samples from our strongest model