Code generation — MATH — MMLU — Code critique

Correctness (y-axis) vs Number of demos (x-axis)

x-axis labels: 0, 1, 8, 64, 512 (Code generation, MATH, Code critique); 0, 1, 64, 512, 4098 (MMLU)

Legend:
- Locked with pwd
- Locked without pwd
- All demos
- Some demos