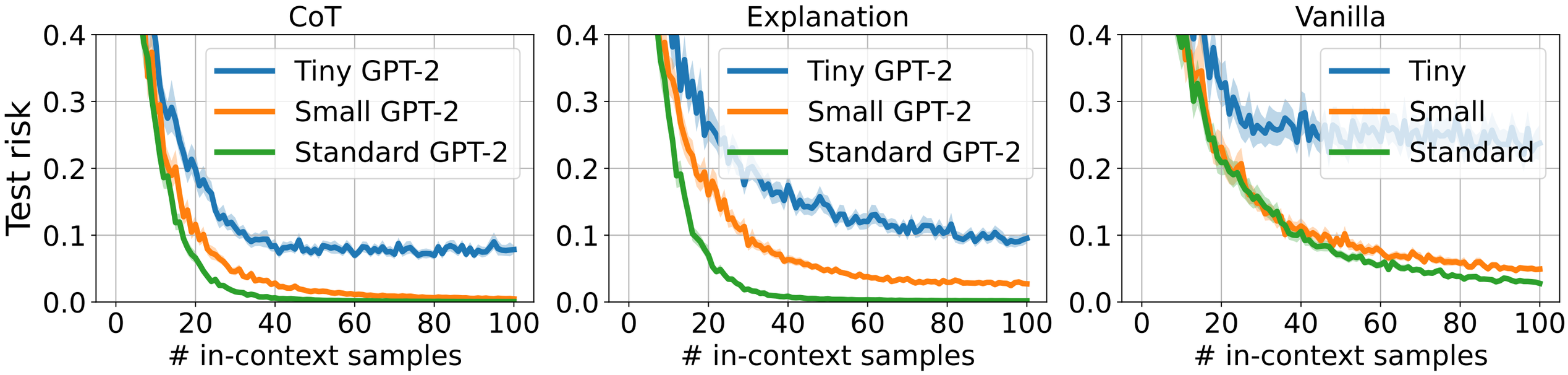
CoT
Explanation
Vanilla
Test risk
0.0
0.1
0.2
0.3
0.4
0
20
40
60
80
100
in-context samples
Tiny GPT-2
Small GPT-2
Standard GPT-2
Tiny
Small
Standard