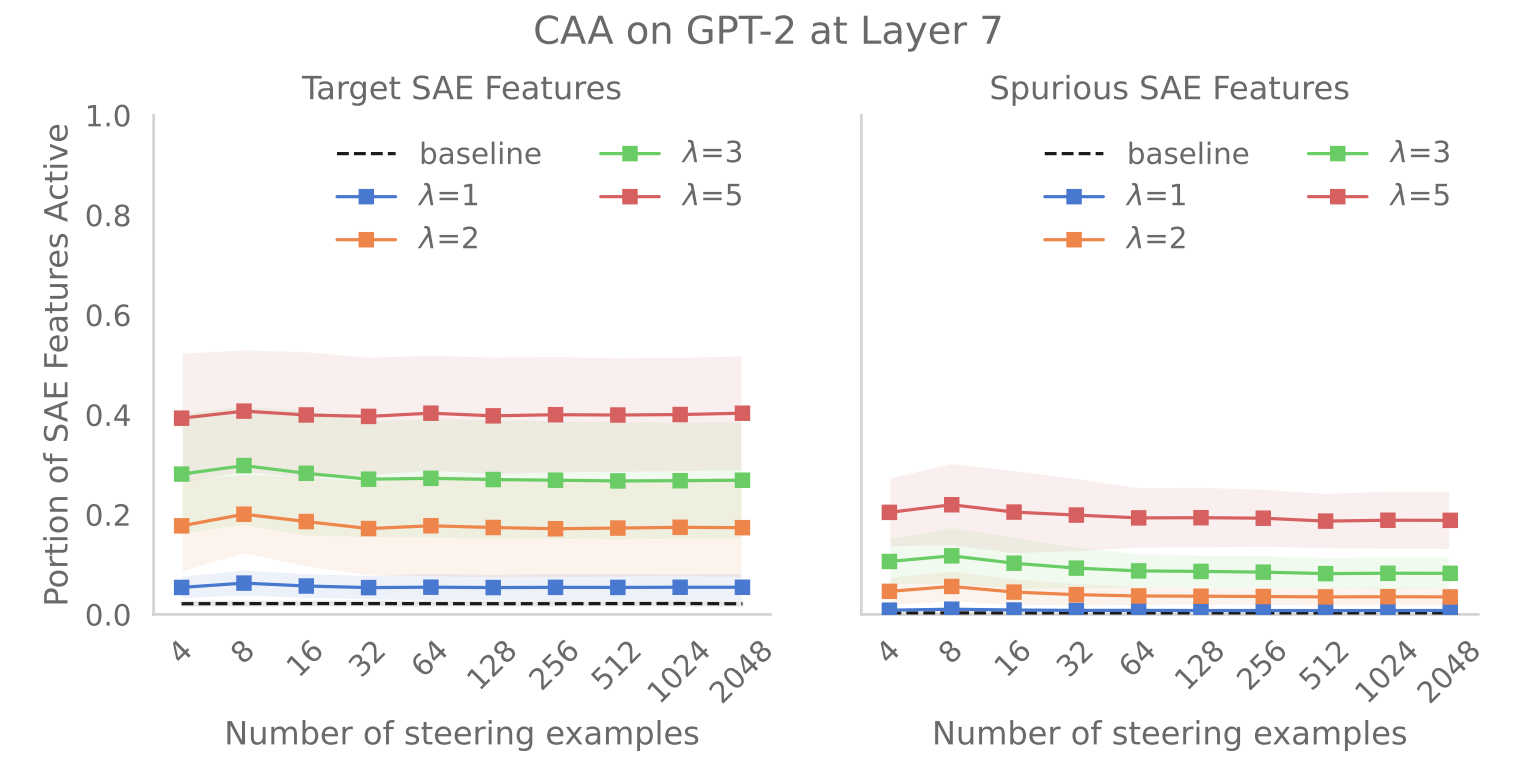
CAA on GPT-2 at Layer 7
Target SAE Features
Spurious SAE Features
Portion of SAE Features Active
1.0
0.8
0.6
0.4
0.2
0.0
baseline
λ=1
λ=2
λ=3
λ=5
4
8
16
32
64
128
256
512
1024
2048
Number of steering examples
Number of steering examples