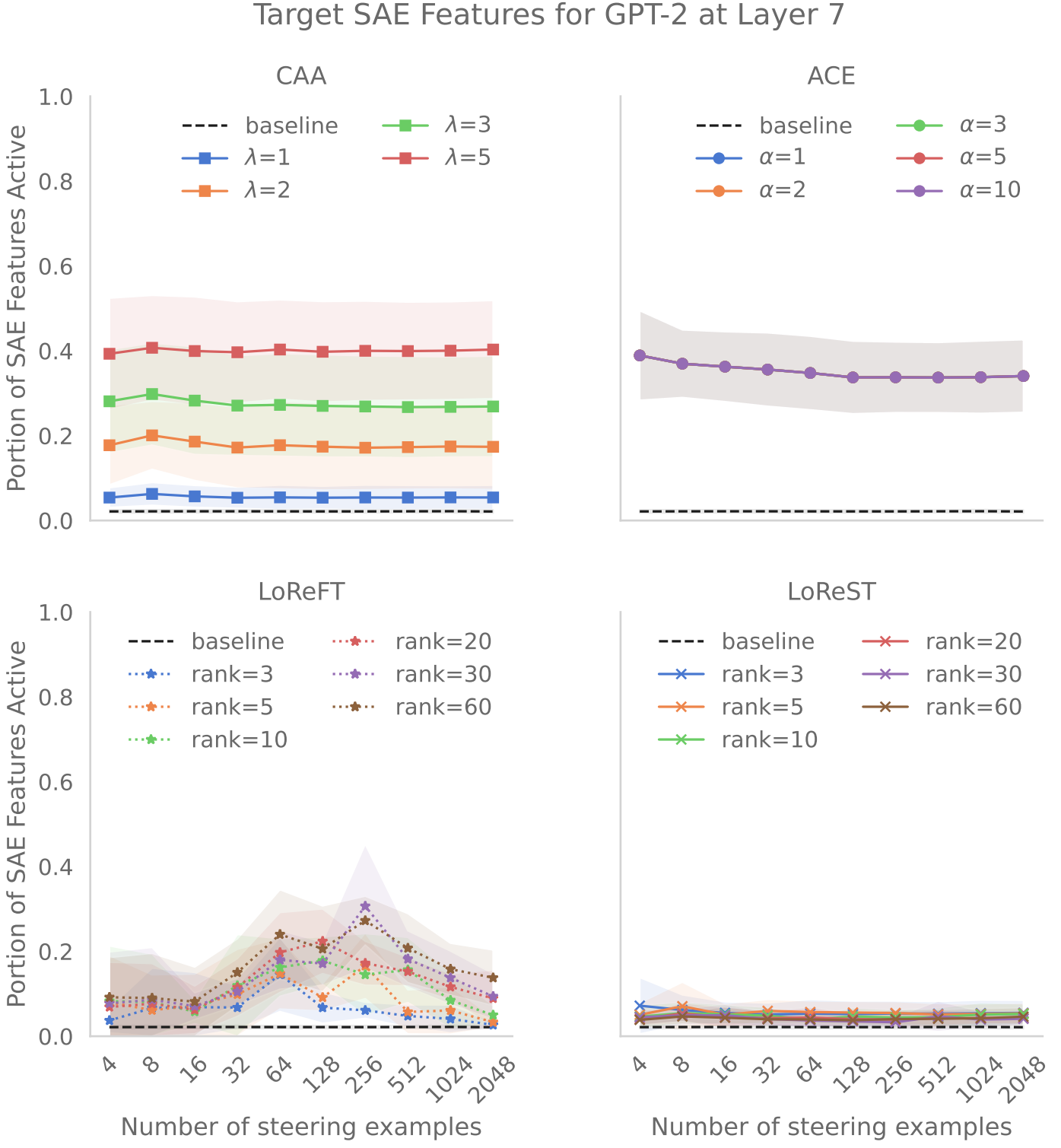
Target SAE Features for GPT-2 at Layer 7
CAA
ACE
LoReFT
LoReST
Portion of SAE Features Active
Number of steering examples
baseline
λ=1
λ=2
λ=3
λ=5
α=1
α=2
α=3
α=5
α=10
rank=3
rank=5
rank=10
rank=20
rank=30
rank=60
0.0
0.2
0.4
0.6
0.8
1.0
4
8
16
32
64
128
256
512
1024
2048