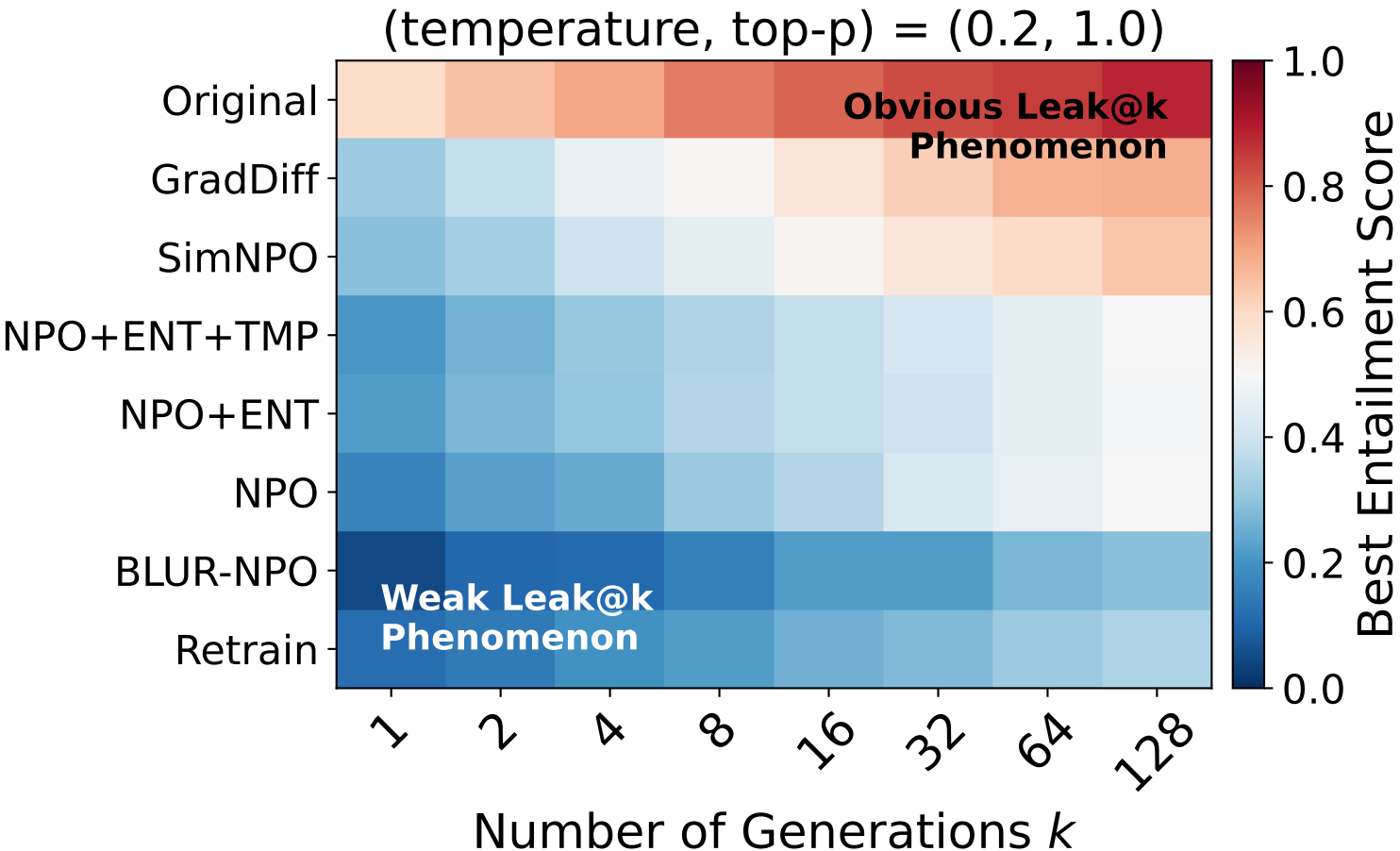
(temperature, top-p) = (0.2, 1.0)
Original
GradDiff
SimNPO
NPO+ENT+TMP
NPO+ENT
NPO
BLUR-NPO
Retrain
Obvious Leak@k Phenomenon
Weak Leak@k Phenomenon
1
2
4
8
16
32
64
128
Number of Generations k
1.0
0.8
0.6
0.4
0.2
0.0
Best Entailment Score