Entailment Score Increment at (temperature, top-p) = (0.2, 1.0)**BLUR-NPO** 0.7 NPO NPO+ENT NPO+ENT+TMP **SimNPO** 0.6 GradDiff Increment of Best Entailment Score 0.5 0.4 0.3 Worst Leak@k Performance 0.2 Best Leak@k Performance 0.1 0.0 -0.1

4

16

Number of Generations (k)

32

64

128