LoReFT on GPT-2 at Layer 7

Target SAE Features | Spurious SAE Features

Legend: baseline, rank=3, rank=5, rank=10, rank=20, rank=30, rank=60

Y-axis: Portion of SAE Features Active

X-axis: Number of steering examples