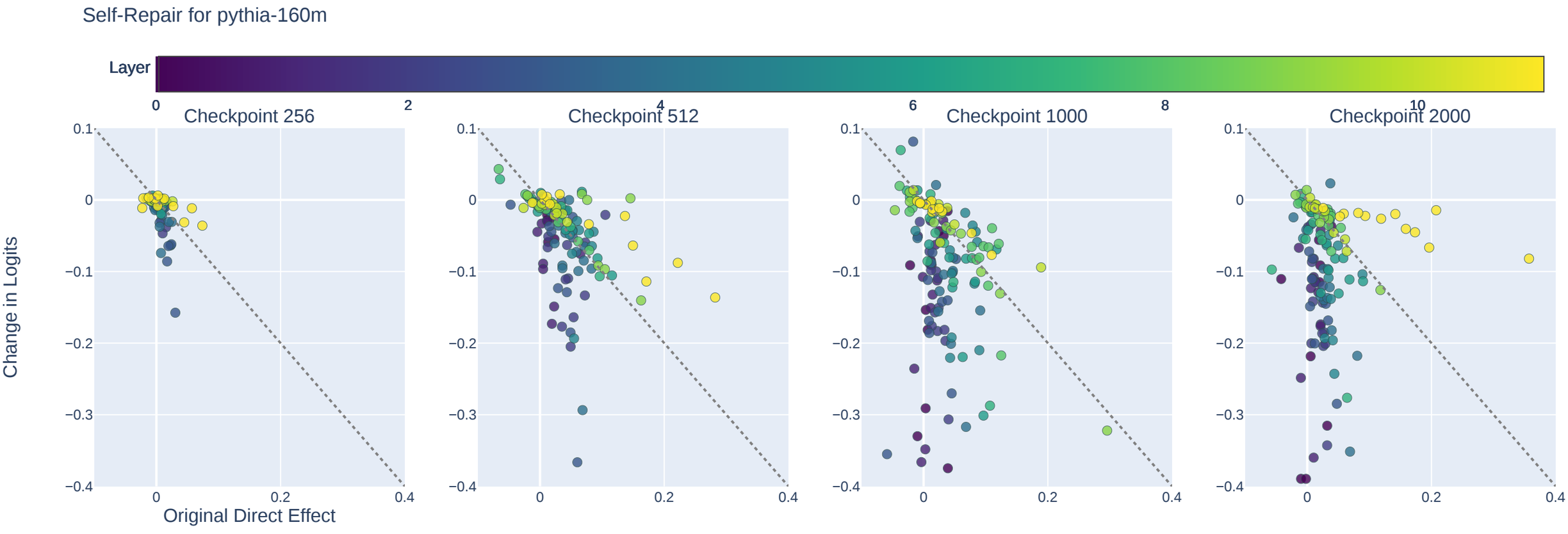
Self-Repair for pythia-160m
Layer
0
2
4
6
8
10
Checkpoint 256
Checkpoint 512
Checkpoint 1000
Checkpoint 2000
Change in Logits
Original Direct Effect
0.1
0
−0.1
−0.2
−0.3
−0.4
0
0.2
0.4