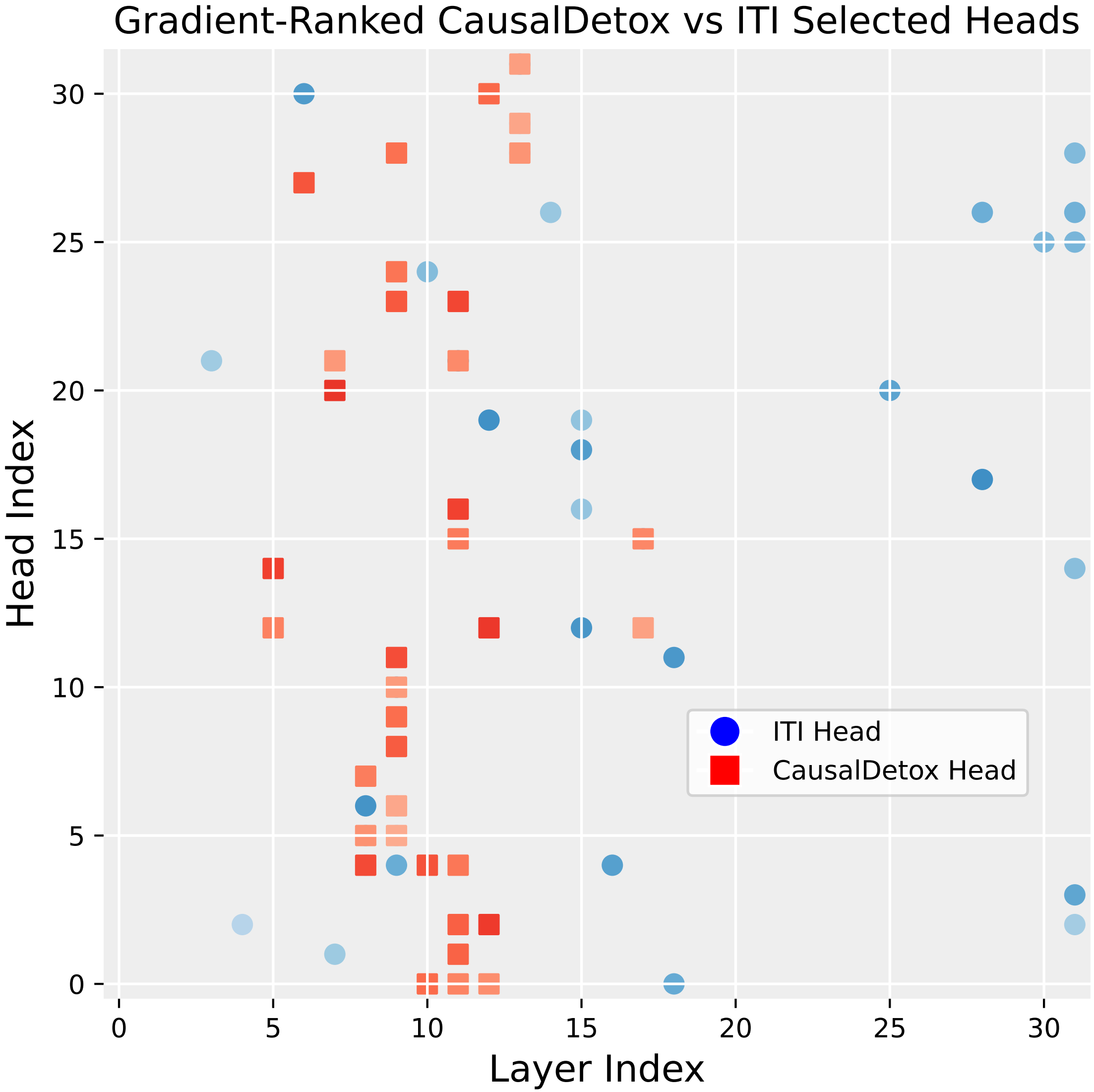
Gradient-Ranked CausalDetox vs ITI Selected Heads
Head Index
0
5
10
15
20
25
30
Layer Index
0
5
10
15
20
25
30
● ITI Head
■ CausalDetox Head