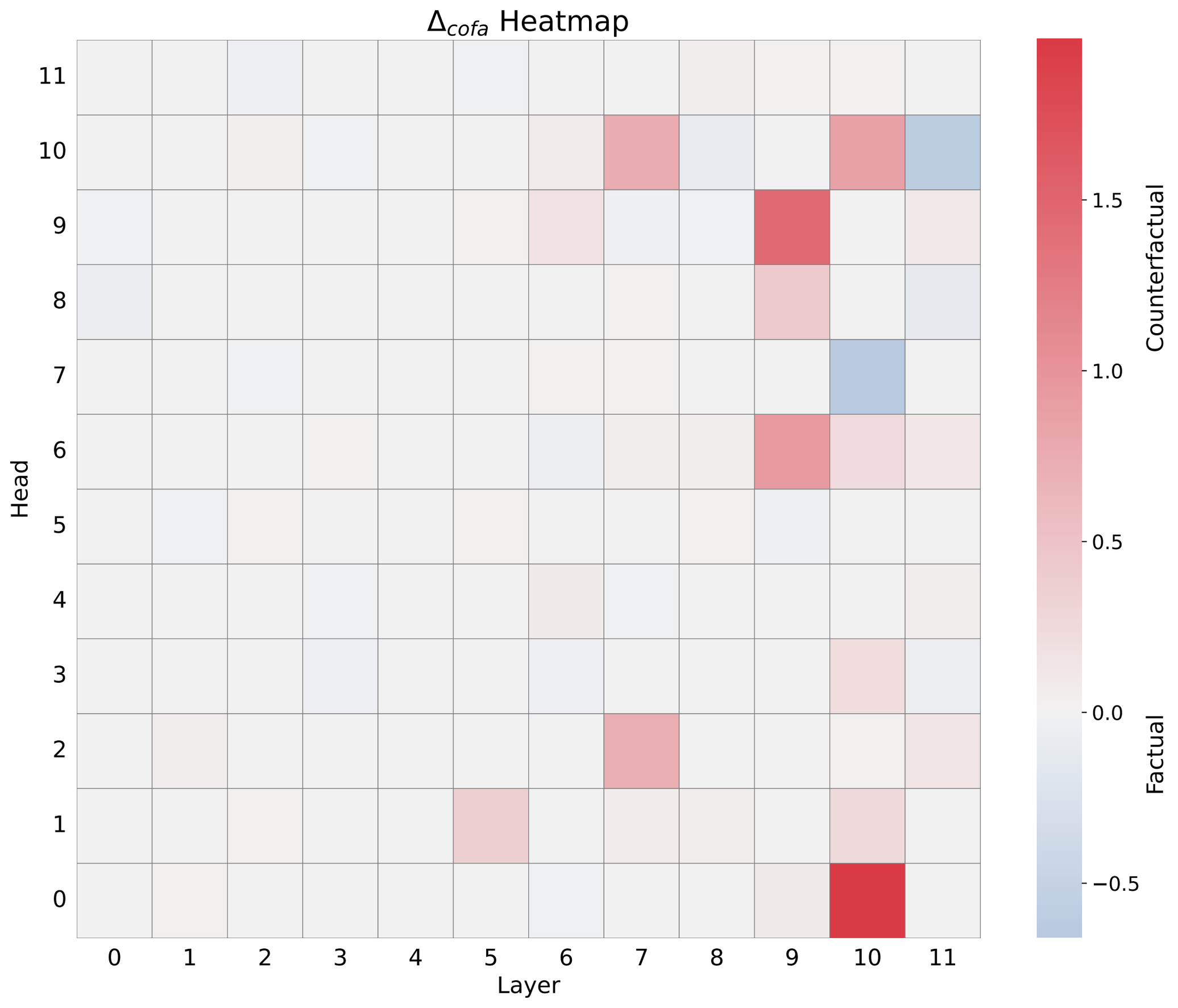
Δ_{cofa} Heatmap
Head
Layer
0
1
2
3
4
5
6
7
8
9
10
11
Counterfactual
Factual
1.5
1.0
0.5
0.0
−0.5