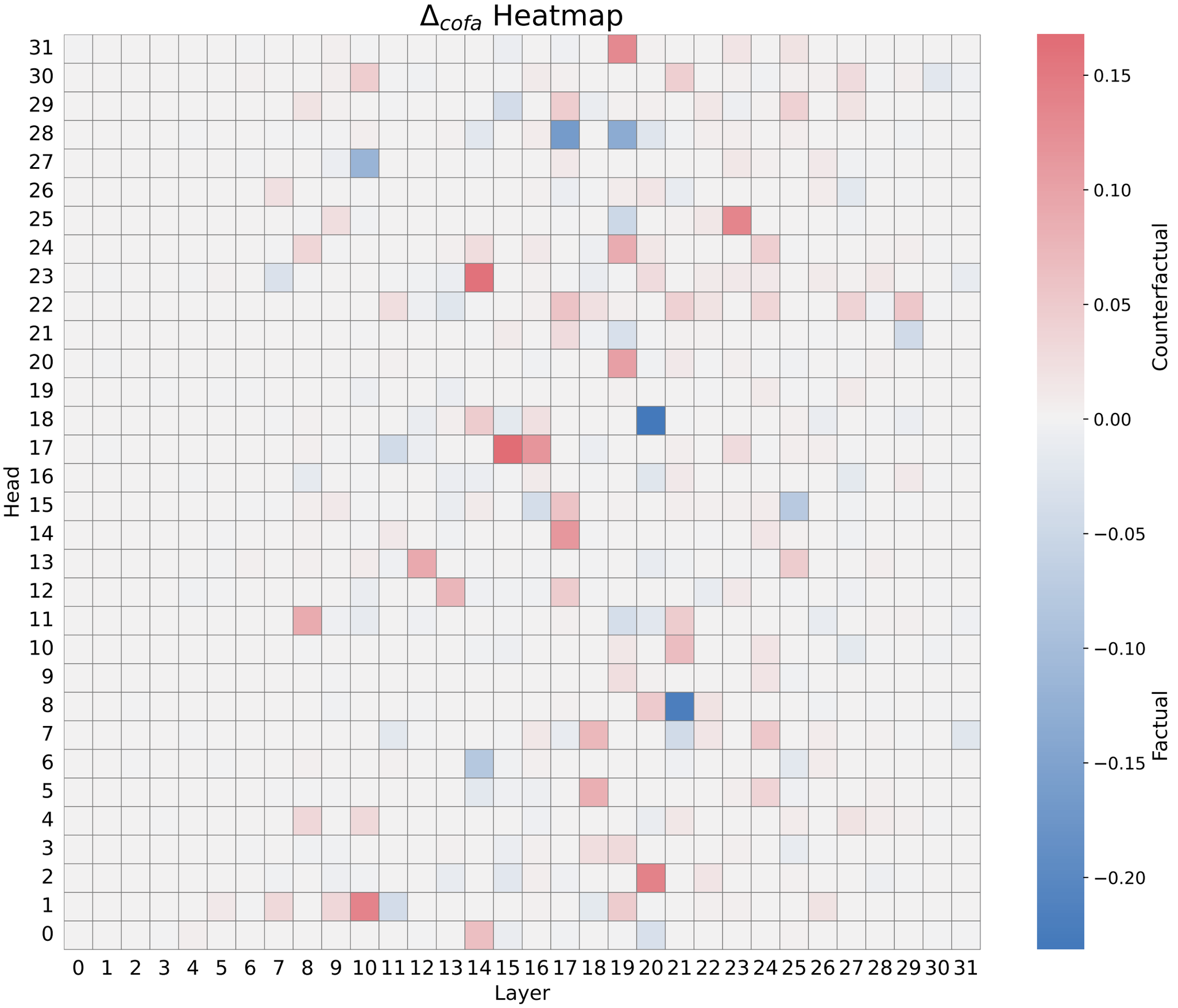
Δ_{cofa} Heatmap
Head
Layer
Counterfactual
Factual
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20