Change of loss after intervention (lower is better), scores v5

Mean ΔLoss

Layer with intervention

Dataset=openwebtext-128-64

Vector multiplier
- -1.0
- -0.5
- -0.2
- -0.15
- -0.1

Loading [MathJax]/extensions/MathMenu.js