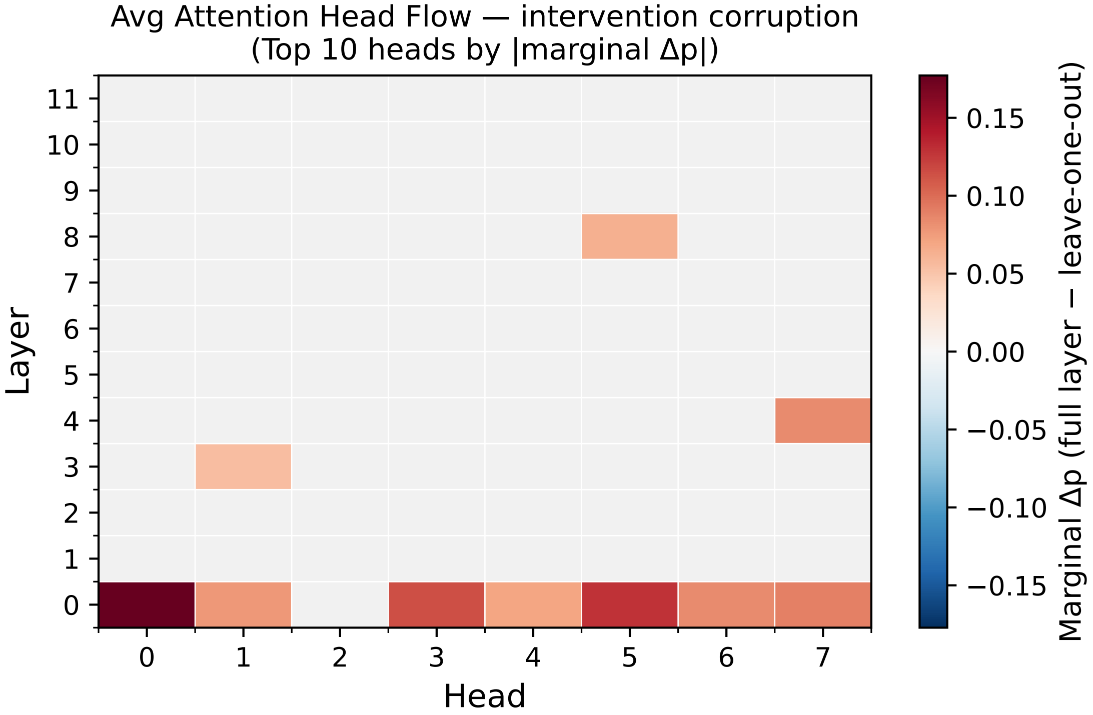
Avg Attention Head Flow — intervention corruption
(Top 10 heads by |marginal Δp|)
Layer
11
10
9
8
7
6
5
4
3
2
1
0
Head
0
1
2
3
4
5
6
7
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
Marginal Δp (full layer − leave-one-out)