

attn_head_out Activation Patching (All Pos) (Top 10 heads by $|\Delta p|$)

