Causal Ablation Effect: Layer × k Sweep

Residual stream layer	k=8	k=16	k=32	k=64
L12	-0.00	-0.01	-0.32	-0.28
L16	-0.03	0.23	0.27	0.43
L20	0.02	0.30	0.32	0.34
L24	0.43	6.30	7.69	8.73

Top-k features ablated

Δ(safe − random) logprob shift