Patching Effects by Layer and Token (Clean: "Paris" vs Corrupted: "Berlin")

Normalized Patching Effect

-0.2

