

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

