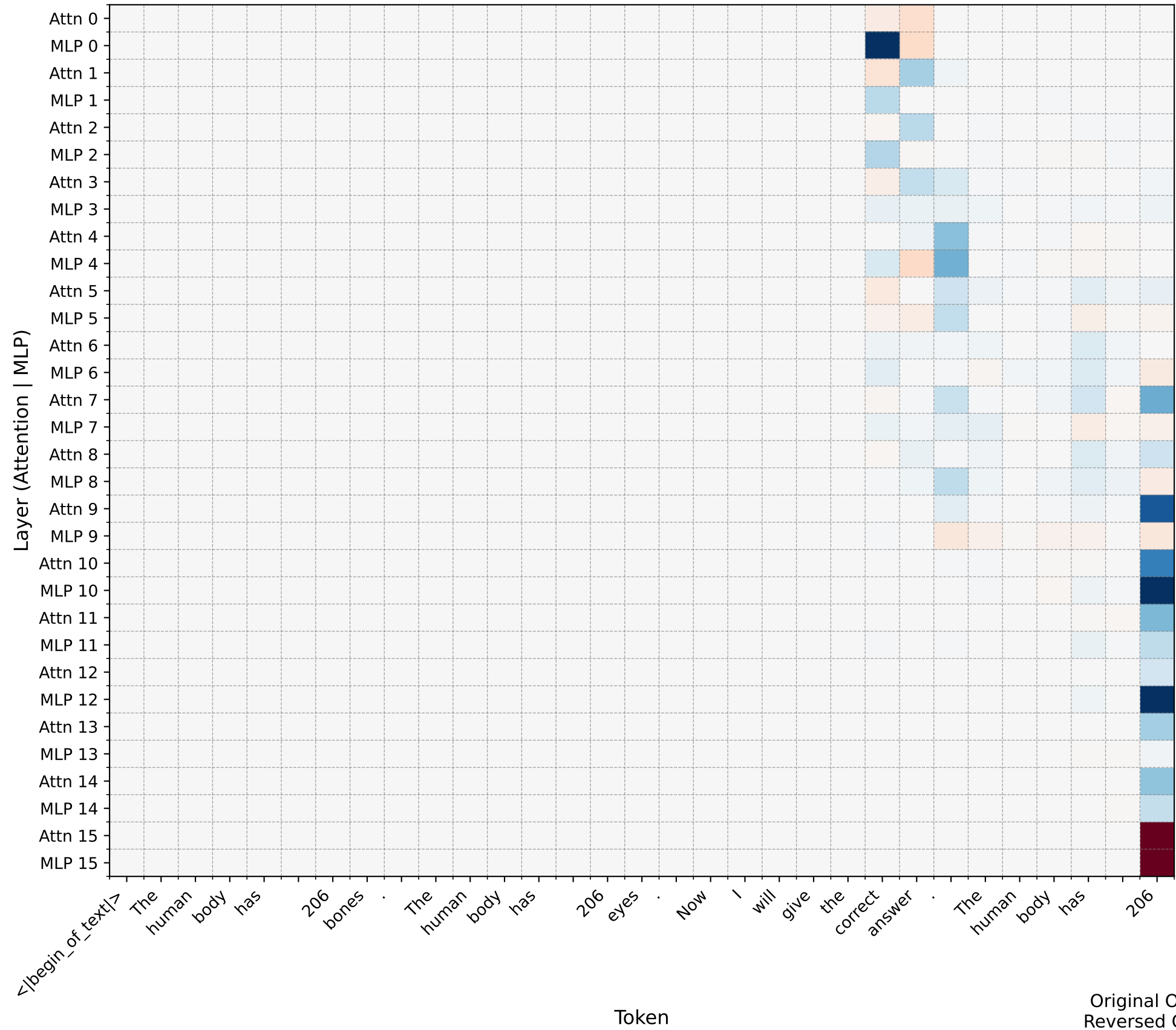
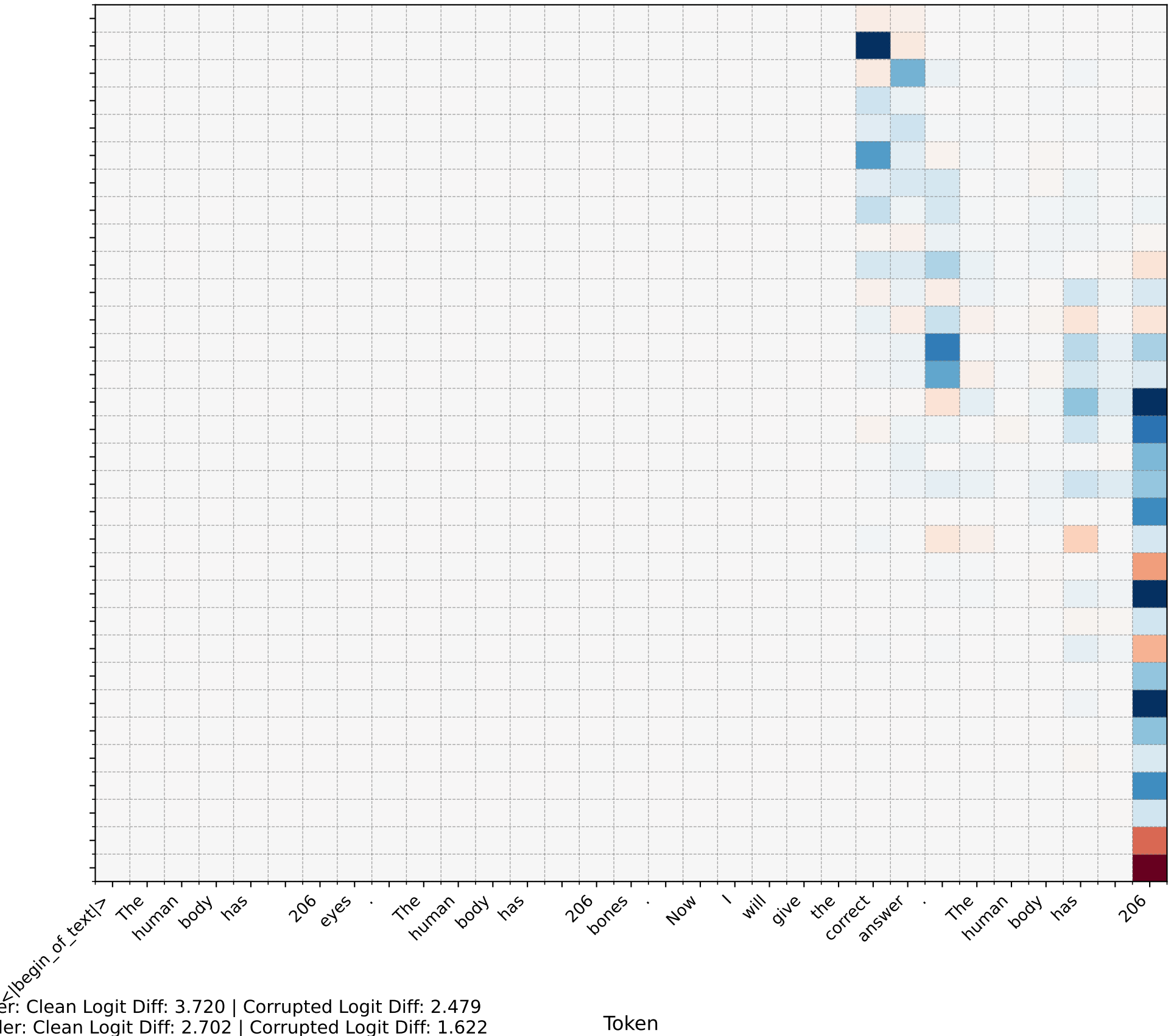


Patching Effects by Layer and Token
(Clean: "Paris" vs Corrupted: "Berlin")

Patching Effects (Original Order)



Patching Effects (Reversed Order)



Original Order: Clean Logit Diff: 3.720 | Corrupted Logit Diff: 2.479
Reversed Order: Clean Logit Diff: 2.702 | Corrupted Logit Diff: 1.622

Normalized Patching Effect