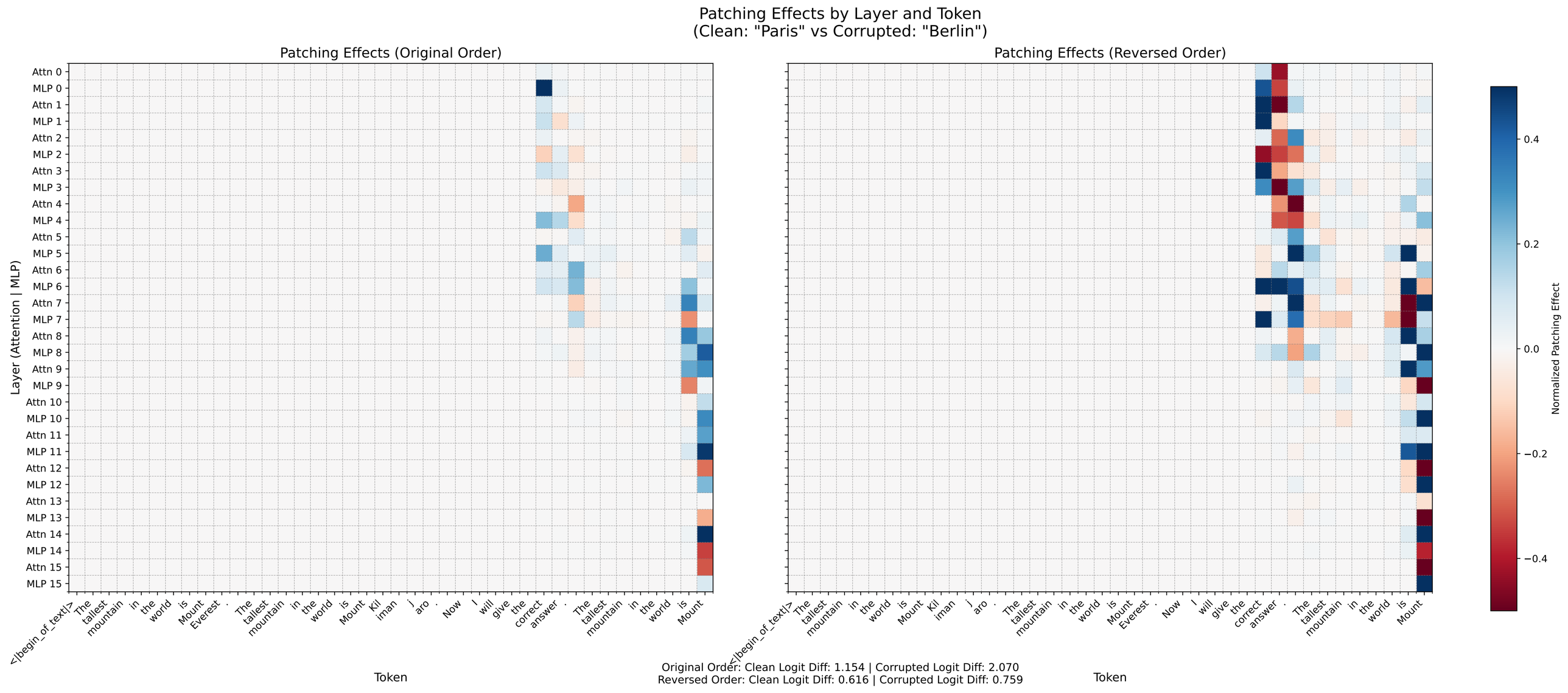
Patching Effects by Layer and Token
(Clean: "Paris" vs Corrupted: "Berlin")
Patching Effects (Original Order)
Patching Effects (Reversed Order)
Layer (Attention | MLP)
Token
Normalized Patching Effect
0.4
0.2
0.0
−0.2
−0.4
Original Order: Clean Logit Diff: 1.154 | Corrupted Logit Diff: 2.070
Reversed Order: Clean Logit Diff: 0.616 | Corrupted Logit Diff: 0.759