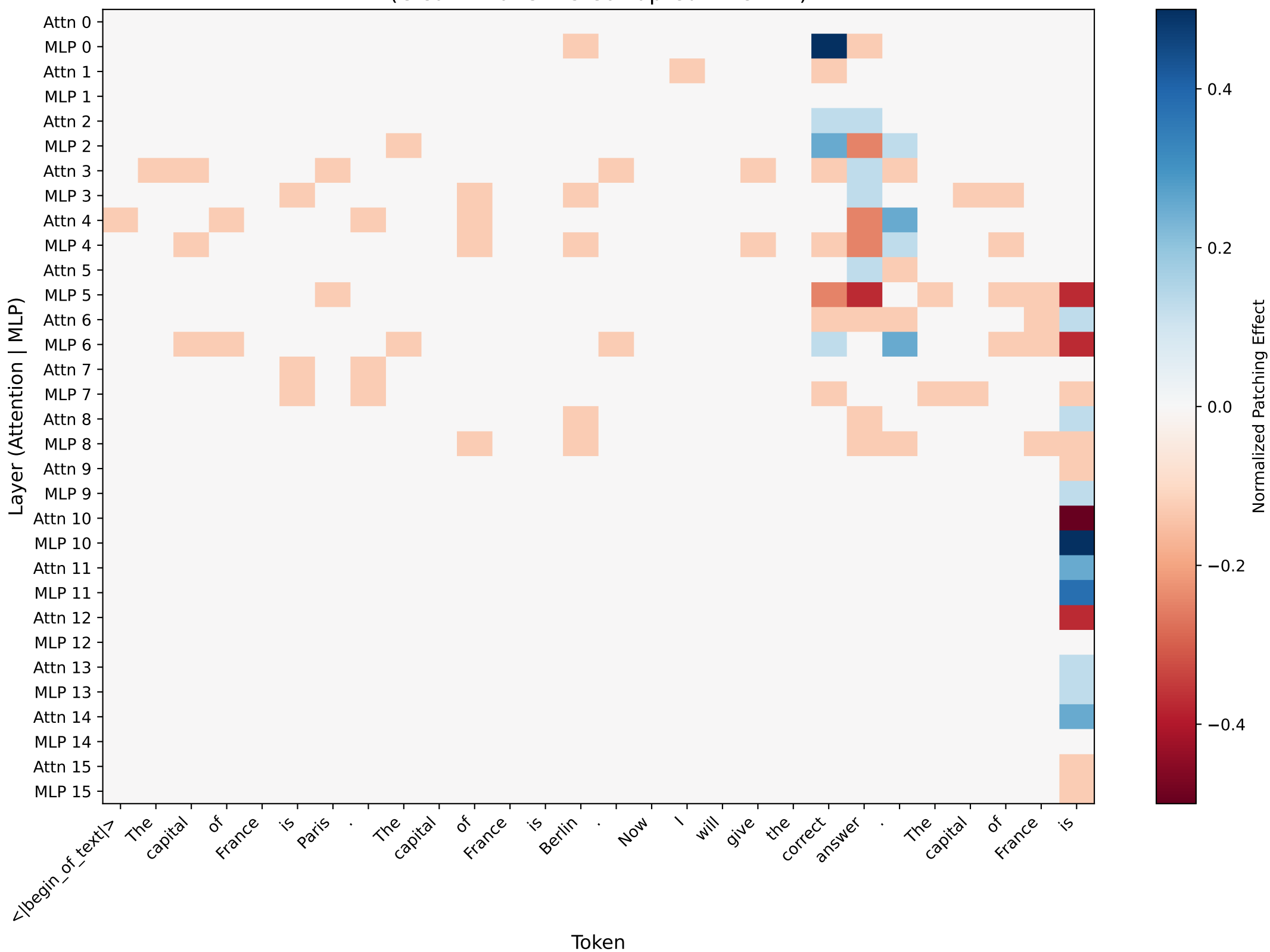


Patching Effects by Layer and Token  
(Clean: "Paris" vs Corrupted: "Berlin")



Clean Logit Diff: 2.875 | Corrupted Logit Diff: 1.875