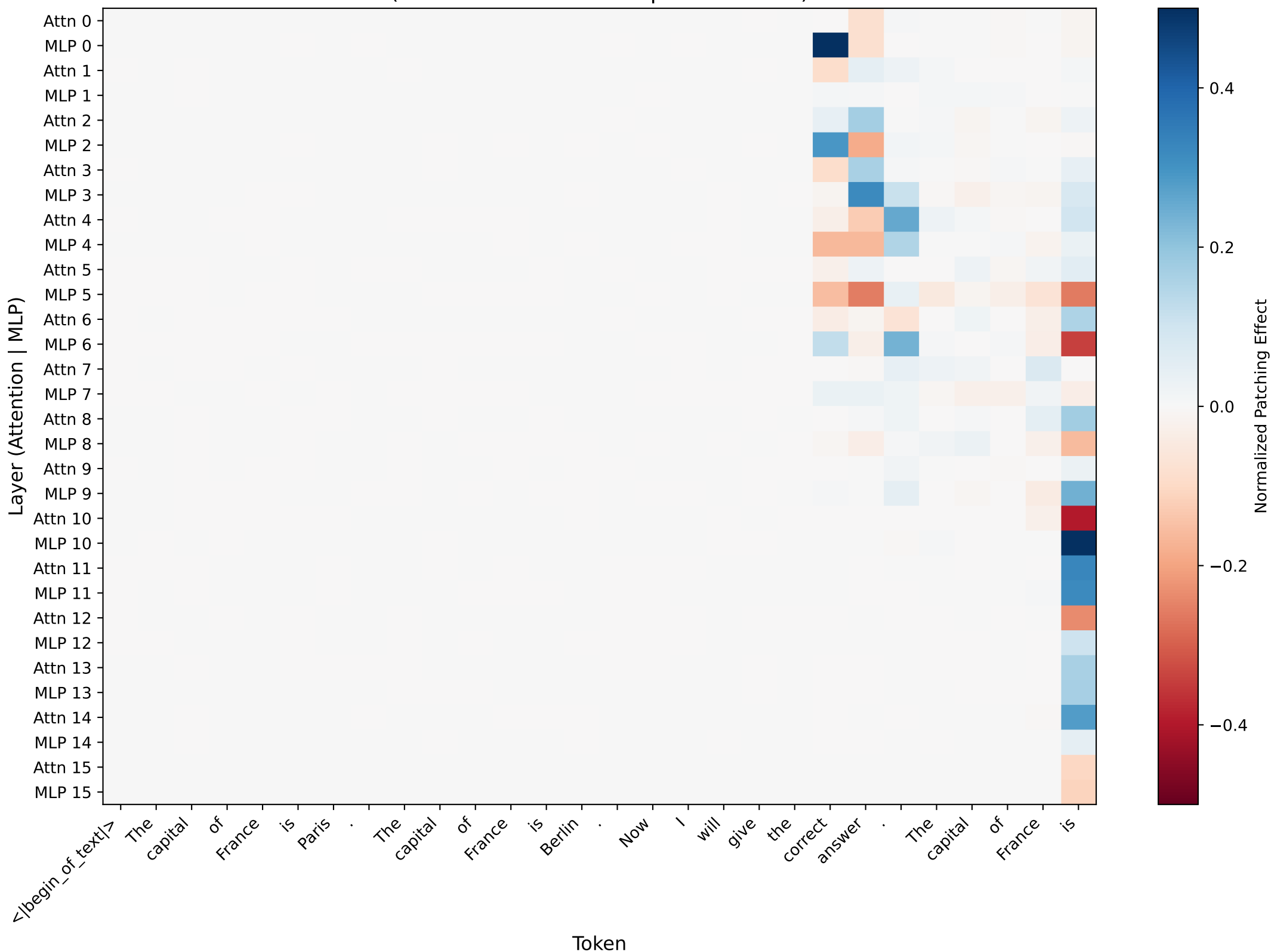


Patching Effects by Layer and Token
(Clean: "Paris" vs Corrupted: "Berlin")



Clean Logit Diff: 2.891 | Corrupted Logit Diff: 1.843