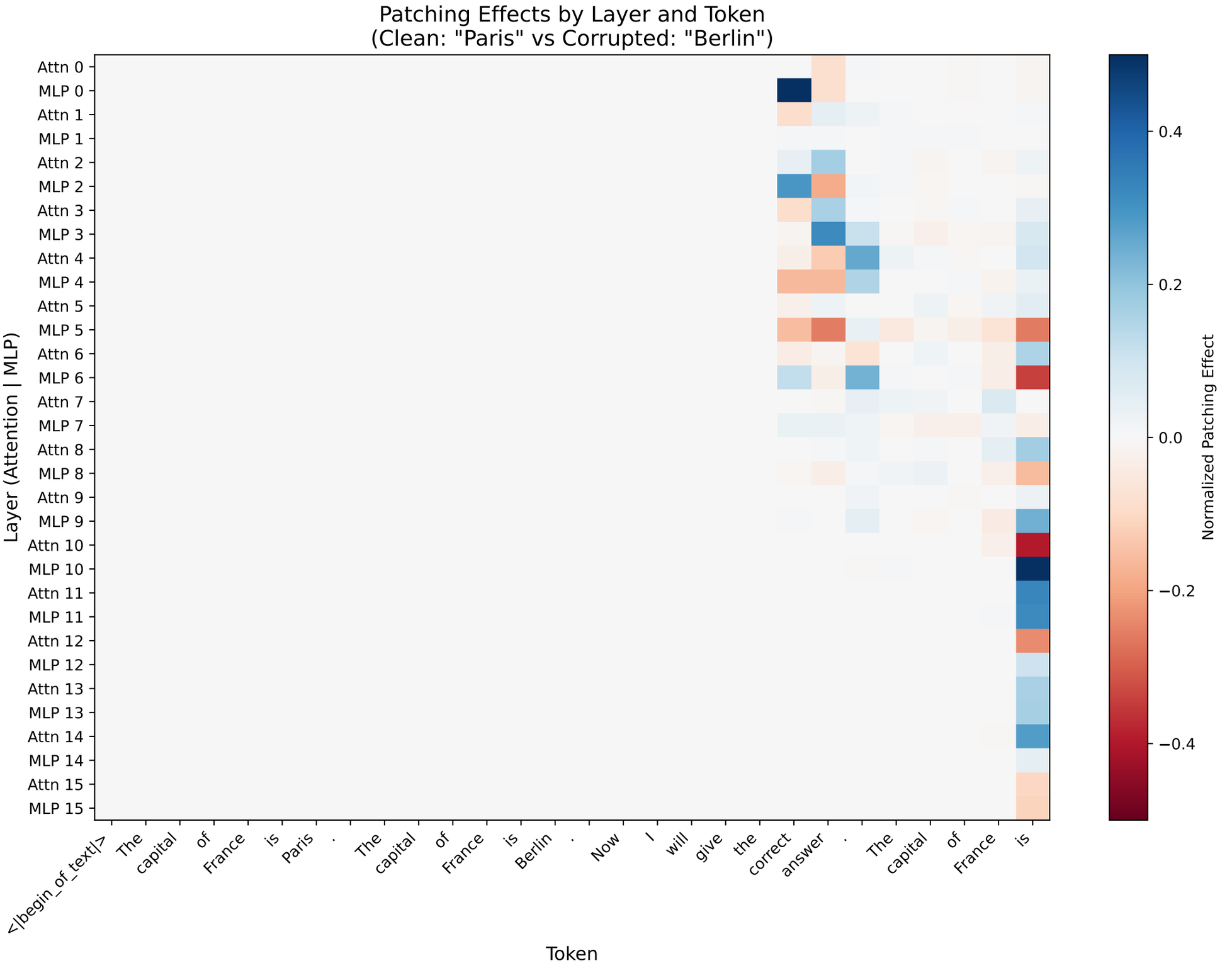


Clean Logit Diff: 2.891 | Corrupted Logit Diff: 1.843