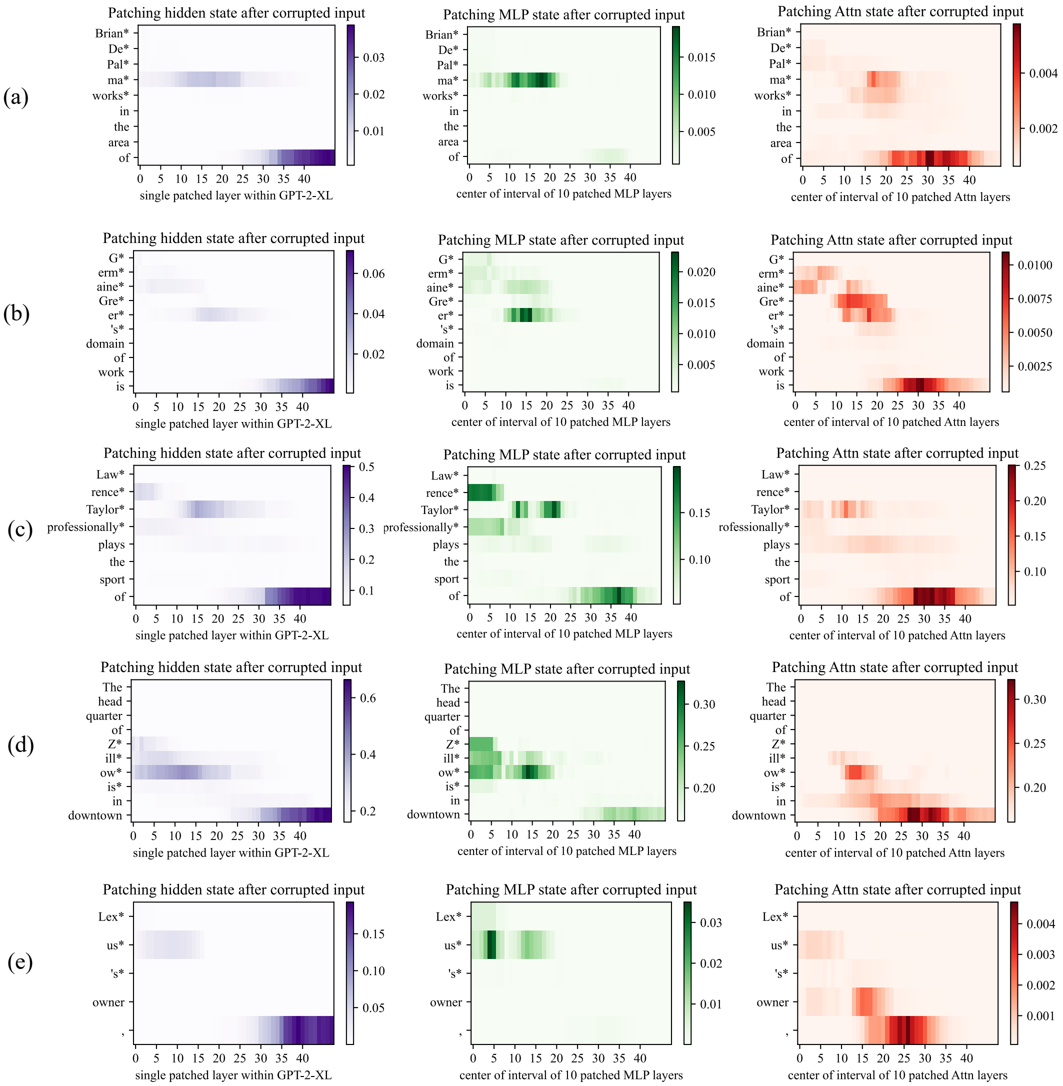

Figure 12: Causal traces in the presence of additional corruption. Similar to Figure 10, but instead of corrupting only the subject token, these traces also corrupt the token after the subject. Causal effects are somewhat reduced due to the the model losing some ability to read the relation between the subject and object, but these traces continue to show concentrated causal effects at the last token of the subject even when the last token is not the last token corrupted. Causal effects of MLP layers at the last subject token continues to be pronounced.