

Figure 11: Causal traces show that the last token of the subject name is not always decisive. (a) shows a typical case: even though the name ‘NTFS’ is a spelled out acronym, the model does MLP lookups at the last letter of the name that are decisive when the model recalls the developer Microsoft. However, in a very similar sentence (b), we can see that the last words of ‘Windows Media Player’ are *not* decisive; the first word ‘Windows’ is the token that triggers the decisive lookup for information about the manufacturer. The information also seems to pass through the attention at the second token ‘Media’. Similarly in (c) we find that the Tokyo headquarters of ‘Mitsubishi Electric’ does not depend on the word ‘Electric’, and in (d) the location of death of Madame de Montesson seems to be mainly determined by the observed title ‘Madame’. In (e) we have a typical low-confidence trace, in which no runs of MLP lookups inside the subject name appear decisive; the model seems to particularly depend on the prompt word ‘performing’ to guess that the subject might play the piano.

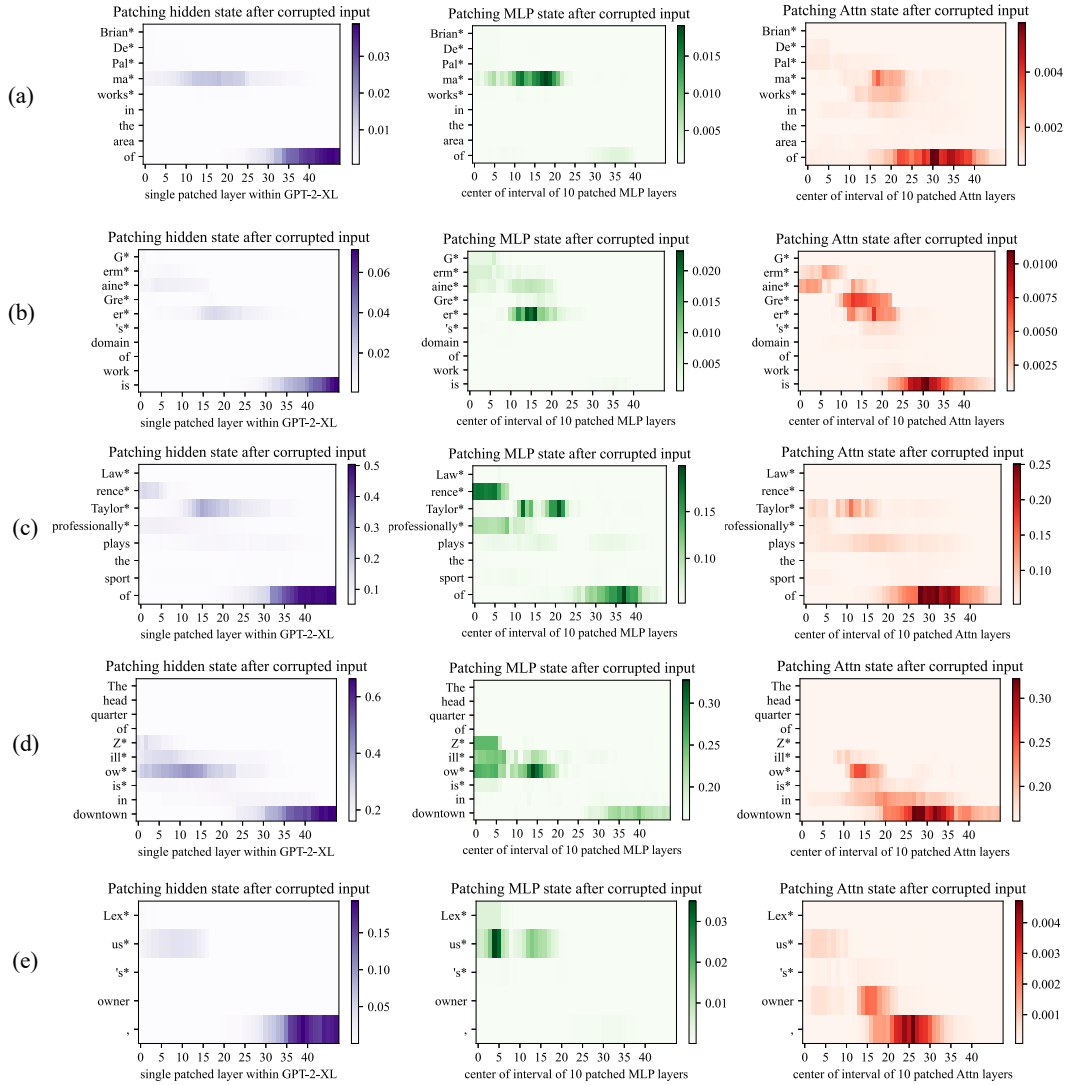


Figure 12: Causal traces in the presence of additional corruption. Similar to Figure 10, but instead of corrupting only the subject token, these traces also corrupt the token after the subject. Causal effects are somewhat reduced due to the the model losing some ability to read the relation between the 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.

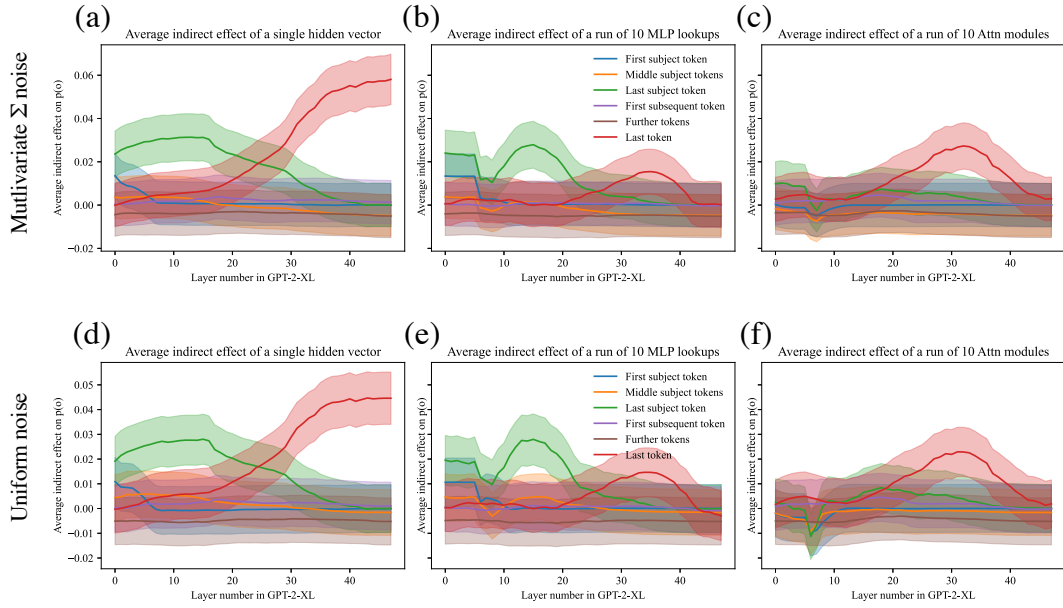


Figure 13: Comparing different noise choices. (Compare to Figure 7, where noise is chosen as a $3\sigma_t$ spherical Gaussian, where σ_t is measured to match the observed spherical variance over tokens.) In a, b, c we draw noise from a multivariate Gaussian $\mathcal{N}(\mu; \Sigma)$ where μ and Σ are chosen to match the observed mean and covariance over a sample of tokens. In d, e, f we draw noise from a uniform distribution in the range $\pm 3\sigma$ instead of a Gaussian distribution. In both cases, the average total effects measured between the clean run and the corrupted run are large enough to measure causal traces, but the effects are smaller than the choice of $3\sigma_t$ used in the main paper.