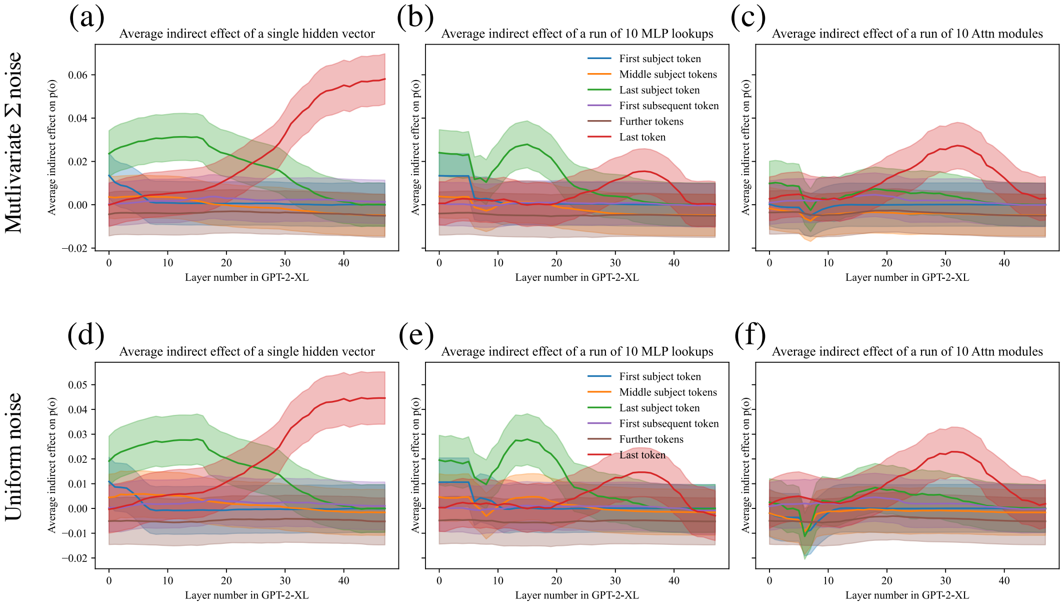

Figure 13: Comparing different noise choices. (Compare to Figure 7, where noise is chosen as a $3\sigma_t$ spherical Gaussian, where σ_t is measured to match the observed spherical variance over tokens.) In a, b, c we we draw noise from a multivariate Gaussian $\mathcal{N}(\mu; \Sigma)$ where μ and Σ 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.