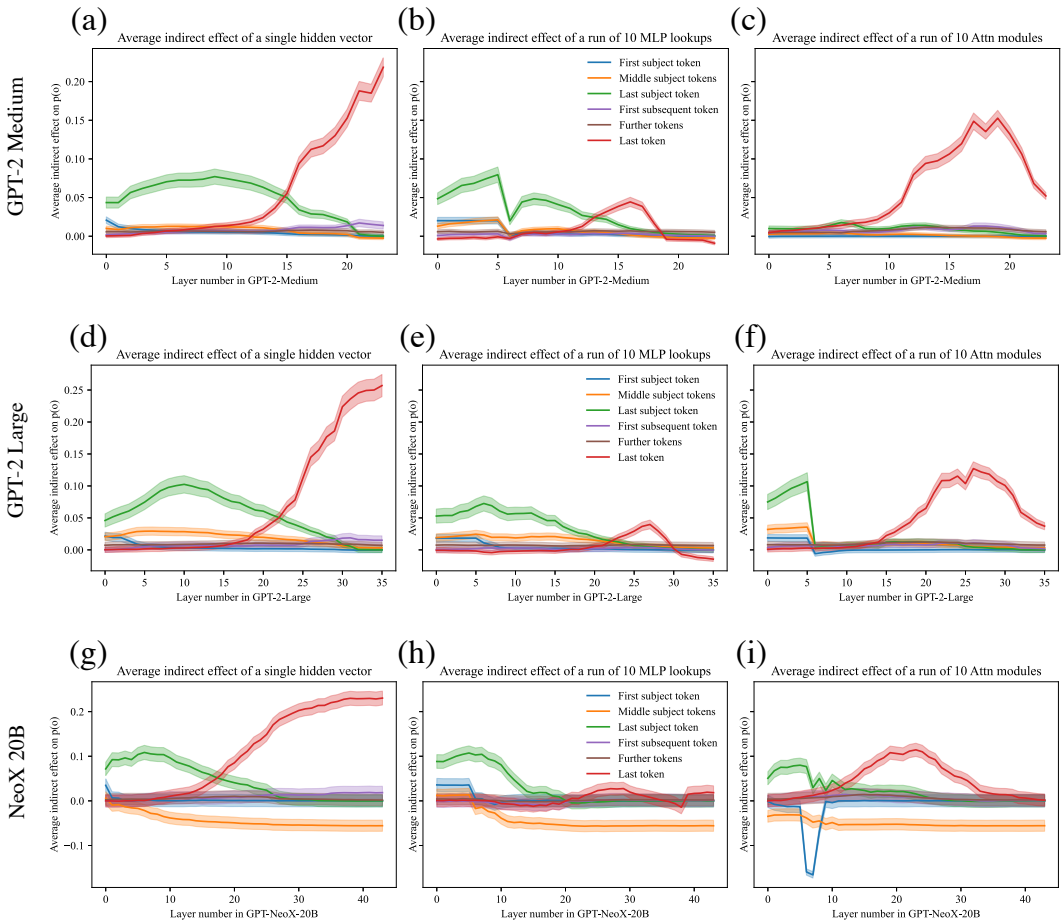

Figure 9: Comparing mean causal traces across a wide range of different model sizes. (Compare to Figure 7.) GPT-medium (a, b, c) has 334 million parameters, GPT-large (d, e, f) has 774 million parameters, and NeoX-20B (g, h, i) has 20 billion parameters. In addition, NeoX has some architectural variations. Despite the wide range of differences, a similar pattern of localized causal effects is seen across models. Interestingly, for very large models, some effects are stronger. For example, hidden states before the last subject token have negative causal effects instead of merely low effects, while hidden states at early layers at the last subject token continue to have large positive effects, continuing to implicate MLP. Also, attention modules with strong causal effects appear earlier in the stack of layers.

where components range in $\pm 3\sigma$, effects large enough for causl tracing but smaller than a Gaussian distribution.

If instead of using spherical Gaussian noise, we draw noise from $\mathcal{N}(mu, \Sigma)$ where we set $\mu = \mu_t$ and $\Sigma_= \Sigma_t$ to match the observed distribution over token embeddings, average total effects are also strong enough to perform causal traces. This is shown in Figure 13.

Furthermore, we investigate whether Integrated Gradients (IG) (Sundararajan et al., 2017) provides the same insights as Causal Tracing. We find that IG is very sensitive to local features but does not yield the same insights about large-scale global logic that we have been able to obtain using causal traces. Figure 16 compares causal traces to IG saliency maps.