B.3 Traces of EleutherAI GPT-NeoX (20B) and GPT-J (6B) and smaller models

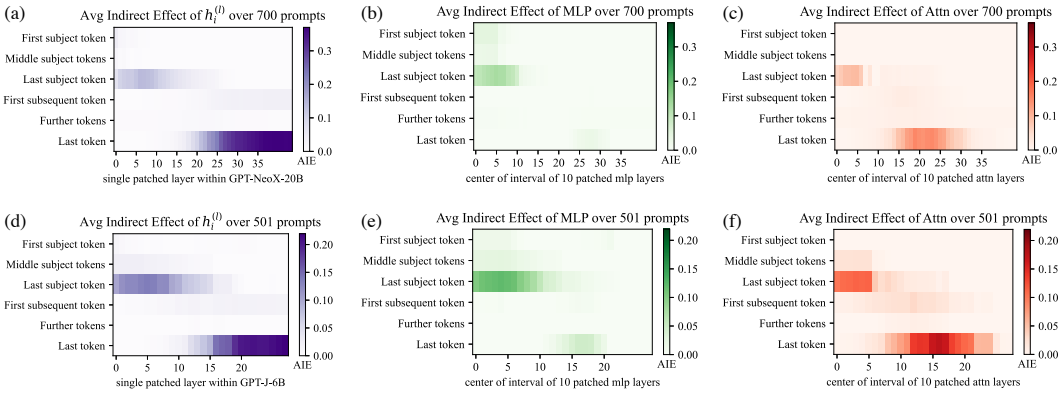

Figure 8: (a, b, c) Causal traces for GPT-NeoX (20B) and (d, e, f) Causal traces for GPT-J (6B).

We conduct the causal trace experiment using on GPT-NeoX (20 billion parameters) as well as GPT-J (6 billion parameters). For GPT-NeoX we adjust the injected noise to $\nu = 0.03$ and in GPT-J we use $\nu = 0.025$ to match embedding magnitudes. We use the same factual prompts as GPT-2 XL, eliminating cases where the larger models would have predicted a different word for the object. Results are shown in Figure 8. GPT-NeoX and GPT-J differ from GPT-2 because they have has fewer layers (44 and 28 layers instead of 48), and a slightly different residual structure across layers. Nevertheless, the causal traces look similar, with an early site with causal states concentrated at the last token of the subject, a large role for MLP states at that site. Again, attention dominates at the last token before prediction.

There are some differences compared to GPT-2. The importance of attention at the first layers of the last subject token is more apparent in GPT-Neo and GPT-J compared to GPT-2, suggesting that the attention parameters may be playing a larger role in storing factual associations. This concentration of attention at the beginning may also be due to fewer layers in the Eleuther models: attending to the subject name must be done in a concentrated way at just a layer or two, because there are not enough layers to spread out that computation in the shallower model. The similarity between the GPT-NeoX and GPT-J and GPT-2 XL traces helps us to understand why ROME continues to work well with higher-parameter models, as seen on our experiments in altering parameters of GPT-J.

To examine effects over a wide range of scales, we also compare causal traces for smaller models GPT-2 Medium and GPT-2 Large. These smaller models are compared to NeoX-20B in Figure 9. We find that across sizes and architectural variations, early-site MLP modules continue to have high indirect causal effects at the last subject token, although the layers where effects peak are different from one model to another.

B.4 Causal Tracing Examples and Further Insights

We include further examples of phenomena that can be observed in causal traces. Figure 10 shows typical examples across different facts. Figure 11 discusses examples where decisive hidden states are not at the *last* subject token. Figure 14 examines traces at an individual token in more detail.

We note that causal tracing depends on a corruption rule to create baseline input for a model that does not contain all the information needed to make a prediction. Therefore we ask: are Causal Tracing results fragile if the exact form of the corruption changes? We test this by expanding the corruption rule: even when additional tokens after the subject name are also corrupted, we find that the results are substantially the same. Figure 12 shows causal traces with the expanded corruption rule. Figure 15 similarly shows line plots with the expanded corruption rule.

We do find that the noise must be large enough to create large average total effects. For example, if noise with variance that is much smaller is used (for example if we set $\sigma = \sigma_t$), average total effects become very small, and the small gap in the behavior between clean runs and corrupted run makes it difficult discern indirect effects of mediators. Similarly, if we use a uniform distribution

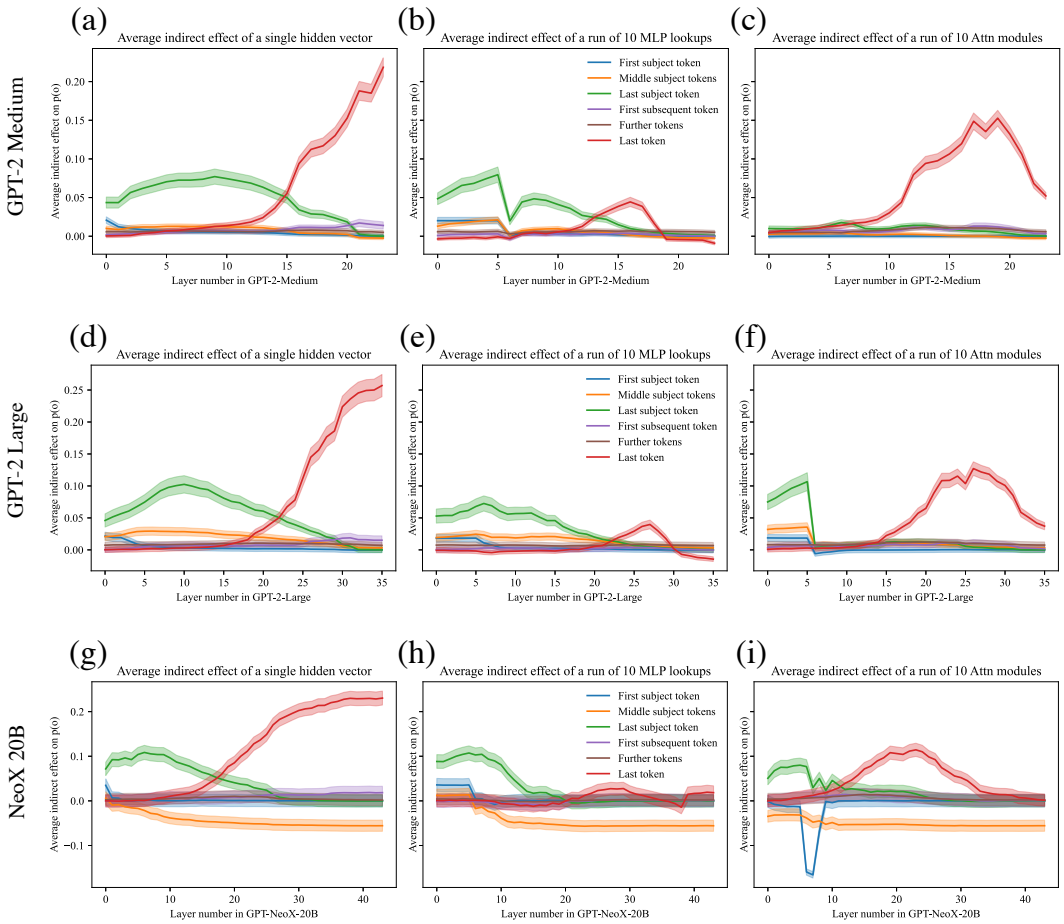

Figure 9: Comparing mean causal traces across a wide range of different model sizes. (Compare to Figure 7.) GPT-medium (a, b, c) has 334 million parameters, GPT-large (d, e, f) has 774 million parameters, and NeoX-20B (g, h, i) has 20 billion parameters. In addition, NeoX has some architectural variations. 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.

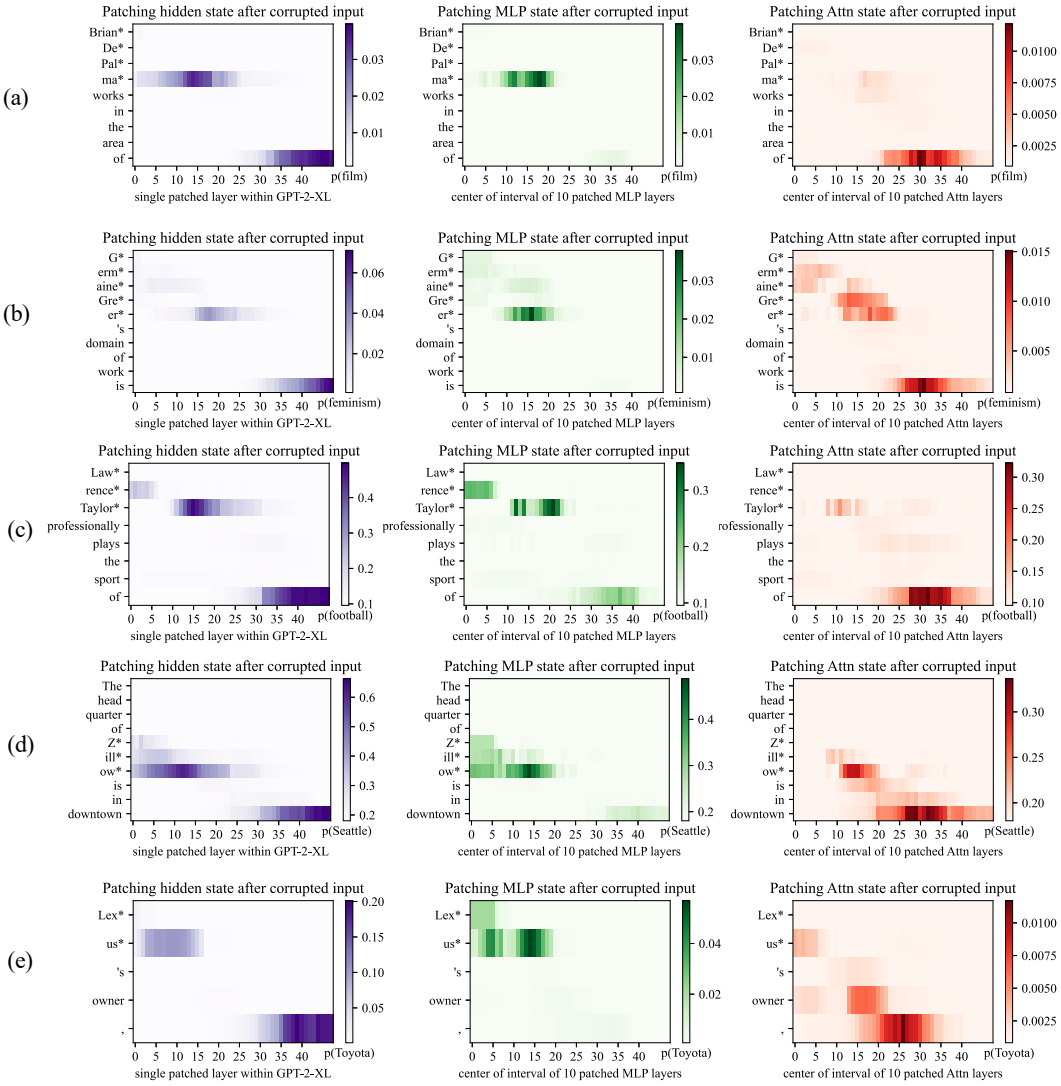

Figure 10: Further examples of causal traces showing appearance of the common lookup pattern on a variety of different types of facts about people and other kinds of entities. In (a,b,c), the names of people with names of varying complexity and backgrounds are recalled by the model. In each case, the MLP lookups on the last token of the name are decisive. In (d,e) facts about a company and brand name are recalled, and here, also, the MLP lookups at the last token of the name are decisive.