B Causal Tracing

B.1 Experimental Settings

Note that, in by-layer experimental results, layers are numbered from 0 to $L-1$ rather than 1 to L.

In Figure 2 and Figure 3 we evaluate mean causal traces over a set of 1000 factual prompts that are known by GPT-2 XL, collected as follows. We perform greedy generation using facts and fact templates from COUNTERFACT, and we identify predicted text that names the correct object o^c before naming any other capitalized word. We use the text up to but not including the object o^c as the prompt, and we randomly sample 1000 of these texts. In this sample of known facts, the predicted probability of the correct object token calculated by GPT-2 XL averages 27.0%.

In the corrupted run, we corrupt the embeddings of the token naming the subject s by adding Gaussian noise $\epsilon \sim \mathcal{N}(0;\nu)$, where $\nu = 3\sigma_t$ is set to be three times larger than the observed standard deviation σ_t of token embeddings as sampled over a body of text. For each run of text, the process is repeated ten times with different samples of corruption noise. On average, this reduces the correct object token score to 8.47%, less than one third the original score.

When we restore hidden states from the original run, we substitute the originally calculated values from the same layer and the same token, and then we allow subsequent calculations to proceed without further intervention. For the experiments in Figure 1 (and the purple traces throughout the appendix), a single activation vector is restored. Naturally, restoring the last vector on the last token will fully restore the original predicted scores, but our plotted results show that there are also earlier activation vectors at a second location that also have a strong causal effect: the average maximum score seen by restoring the most impactful activation vector at the last token of the subject is 19.5%. In Figure 1j where effects are bucketed by layer, the maximum effect is seen around the 15th layer of the last subject token, where the score is raised on average to 15.0%.

B.2 Separating MLP and Attn Effects

When decomposing the effects into MLP and Attn lookups, we found that restoring single activation vectors from individual MLP and individual Attn lookups had generally negligible effects, suggesting the decisive information is accumulated across layers. Therefore for MLP and Attn lookups, we restored runs of ten values of $m_i^{(l)}$ (and $a_i^{(l)}$, respectively) for an interval of layers ranging from $[l_* - 4, ..., l_* + 5]$ (clipping at the edges), where the results are plotted at layer l_*. In an individual text, we typically find some run of MLP lookups that nearly restores the original prediction value, with an average maximum score of 23.6%. Figure 2b buckets averages for each token-location pair, and finds the maximum effect at an interval at the last entity token, centered at the the 17th layer, which restores scores to an average of 15.0%. For Attn lookups (Figure 2c), the average maximum score over any location is 19.4%, and when bucketed by location, the maximum effect is centered at the 32nd layer at the last word before prediction, which restores scores to an average of 16.5%.

Figure 7 shows mean causal traces as line plots with 95% confidence intervals, instead of heatmaps.

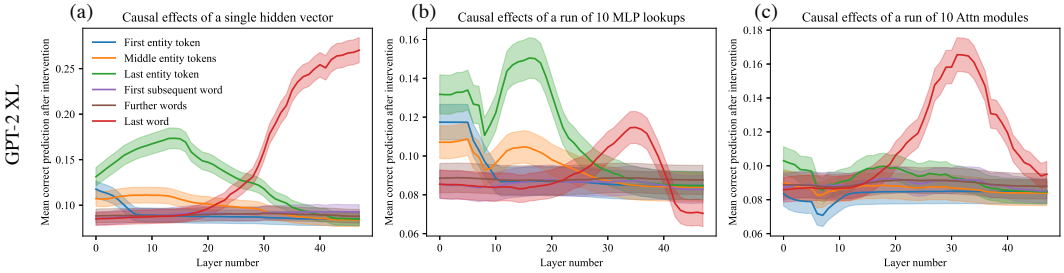

Figure 7: Mean causal traces of GPT-XL over a sample of 1000 factual statements, shown as a line plot with 95% confidence intervals. (a) Shows the same data as Figure 1j as a line plot instead of a heatmap; (b) matches Figure 1k; (c) matches Figure 1m. The confidence intervals confirm that the distinctions between peak and non-peak causal effects at both early and late sites are significant.