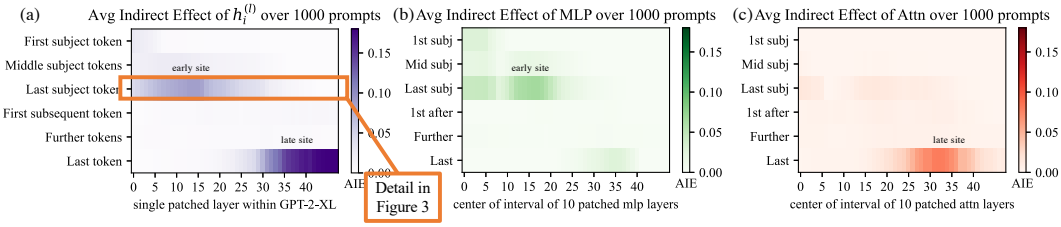

Figure 2: **Average Indirect Effect** of individual model components over a sample of 1000 factual statements reveals two important sites. (a) Strong causality at a 'late site' in the last layers at the last token is unsurprising, but strongly causal states at an 'early site' in middle layers at the last subject token is a new discovery. (b) MLP contributions dominate the early site. (c) Attention is important at the late site. Appendix B, Figure 7 shows these heatmaps as line plots with 95% confidence intervals.

Each layer's MLP is a two-layer neural network parameterized by matrices $W_{proj}^{(l)}$ and $W_{fc}^{(l)}$, with rectifying nonlinearity σ and normalizing nonlinearity γ. For further background on transformers, we refer to Vaswani et al. (2017).[3]

2.1 Causal Tracing of Factual Associations

The grid of states (Figure 1) forms a *causal graph* (Pearl, 2009) describing dependencies between the hidden variables. This graph contains many paths from inputs on the left to the output (next-word prediction) at the lower-right, and we wish to understand if there are specific hidden state variables that are more important than others when recalling a fact.

As Vig et al. (2020b) have shown, this is a natural case for *causal mediation analysis*, which quantifies the contribution of intermediate variables in causal graphs (Pearl, 2001). To calculate each state's contribution towards a correct factual prediction, we observe all of G's internal activations during three runs: a **clean** run that predicts the fact, a **corrupted** run where the prediction is damaged, and a **corrupted-with-restoration** run that tests the ability of a single state to restore the prediction.

- In the **clean run**, we pass a factual prompt x into G and collect all hidden activations $\{h_i^{(l)} \mid i \in [1,T], l \in [1,L]\}$. Figure 1a provides an example illustration with the prompt: "The Space Needle is in downtown _____", for which the expected completion is $o =$ "Seattle".
- In the baseline **corrupted run**, the subject is obfuscated from G before the network runs. Concretely, immediately after x is embedded as $[h_1^{(0)}, h_2^{(0)}, \ldots, h_T^{(0)}]$, we set $h_i^{(0)} := h_i^{(0)} + \epsilon$ for all indices i that correspond to the subject entity, where $\epsilon \sim \mathcal{N}(0; \nu)$[4]; . G is then allowed to continue normally, giving us a set of corrupted activations $\{h_{i*}^{(l)} \mid i \in [1,T], l \in [1,L]\}$. Because G loses some information about the subject, it will likely return an incorrect answer (Figure 1b).
- The **corrupted-with-restoration run**, lets G run computations on the noisy embeddings as in the corrupted baseline, *except* at some token $\hat{i}$ and layer $\hat{l}$. There, we hook G so that it is forced to output the clean state $h_{\hat{i}}^{(\hat{l})}$; future computations execute without further intervention. Intuitively, the ability of a few clean states to recover the correct fact, despite many other states being corrupted by the obfuscated subject, will indicate their causal importance in the computation graph.

Let $\mathbb{P}[o]$, $\mathbb{P}_*[o]$, and $\mathbb{P}_{*,\,\text{clean } h_i^{(l)}}[o]$ denote the probability of emitting o under the clean, corrupted, and corrupted-with-restoration runs, respectively; dependence on the input x is omitted for notational simplicity. The **total effect** (TE) is the difference between these quantities: $\text{TE} = \mathbb{P}[o] - \mathbb{P}_*[o]$. The **indirect effect** (IE) of a specific mediating state $h_i^{(l)}$ is defined as the difference between the probability of o under the corrupted version and the probability when that state is set to its clean version, while the subject remains corrupted: $\text{IE} = \mathbb{P}_{*,\,\text{clean } h_i^{(l)}}[o] - \mathbb{P}_*[o]$. Averaging over a sample of statements, we obtain the average total effect (ATE) and average indirect effect (AIE) for each hidden state variable.[5]

[3] Eqn. 1 calculates attention sequentially after the MLP module as in Brown et al. (2020). Our methods also apply to GPT variants such as Wang & Komatsuzaki (2021) that put attention in parallel to the MLP.

[4] We select ν to be 3 times larger than the empirical standard deviation of embeddings; see Appendix B.1 for details, and see Appendix B.4 for an analysis of other corruption rules.

[5] One could also compute the direct effect, which flows through other model components besides the chosen mediator. However, we found this effect to be noisy and uninformative, in line with results by Vig et al. (2020b).