J Correlation Analysis of Causal Subspaces and Attention Heads

This section identifies the attention heads that align with the causal subspaces discovered in the previous sections. Specifically, first we focus on attention heads whose query projections are aligned with the subspaces—characterized by the relevant singular vectors—that contain the correct answer state OI. To quantify this alignment between attention heads and causal subspaces, we use the following computation.

Let $Q \in \mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}$ denote the query projection weight matrix for a given layer:

We normalize Q column-wise:

$$\tilde{Q}_{:,j} = \frac{Q_{:,j}}{\|Q_{:,j}\|} \quad \text{for each column } j \tag{7}$$

Let $S \in \mathbb{R}^{d_{\text{model}} \times k}$ represent the matrix of k singular vectors (i.e., the causal subspace basis). We project the normalized query weights onto this subspace:

$$Q_{\text{sv}} = \tilde{Q} \cdot S \tag{8}$$

We then reshape the resulting projection into per-head components. Assuming $Q_{\text{sv}} \in \mathbb{R}^{d_{\text{model}} \times k}$, and each attention head has dimensionality d_h, we write:

$$Q_{\text{head}}^{(i)} = Q_{\text{sv}}^{(i)} \in \mathbb{R}^{d_h \times k} \quad \text{for } i = 1, \ldots, n_{\text{heads}} \tag{9}$$

Finally, we compute the norm of each attention head's projection:

$$\text{head\_norm}_i = \left\| Q_{\text{head}}^{(i)} \right\|_F \quad \text{for } i = 1, \ldots, n_{\text{heads}} \tag{10}$$

We compute the $head\_norm$ for each attention head in every layer, which quantifies how strongly a given head reads from the causal subspace present in the residual stream. The results are presented in Fig. 19, and they align with our previous findings: attention heads in the later layers form the QK-circuit by using pointer and address information to retrieve the payload during the Answer lookback.

We perform a similar analysis to check which attention heads' value projection matrix align with the causal subspace that encodes the payload of the Answer lookback. Results are shown in Fig. 20, indicating that attention heads at later layers primarily align with causal subspace containing the answer token.

K Belief Tracking Mechanism in BigToM Benchmark

This section presents preliminary evidence that the mechanisms outlined in Sections 5 and 6 generalize to other benchmark datasets. Specifically, we demonstrate that Llama-3-70B-Instruct answers the belief questions (true belief and false belief) in the BigToM dataset Gandhi et al. (2024) in a manner similar to that observed for CausalToM: by first converting token values to their corresponding OIs and then performing logical operations on them using lookbacks. However, as noted in Section 3, BigToM—like other benchmarks—lacks the coherent structure necessary for causal analysis. As a result, we were unable to replicate all experiments conducted on CausalToM. Thus, the results reported here provide only preliminary evidence of a similar underlying mechanism.

To justify the presence of OIs, we conduct an interchange intervention experiment, similar to the one described in Section H, aiming to localize the character OI at the character token in the question sentence. We construct an original sample by replacing its question sentence with that of a counterfactual sample, selected directly from the unaltered BigToM dataset. Consequently, when processing the original sample, the model has no information about the queried character and, as a result, produces unknown as the final output. However, if we replace the residual vector at the

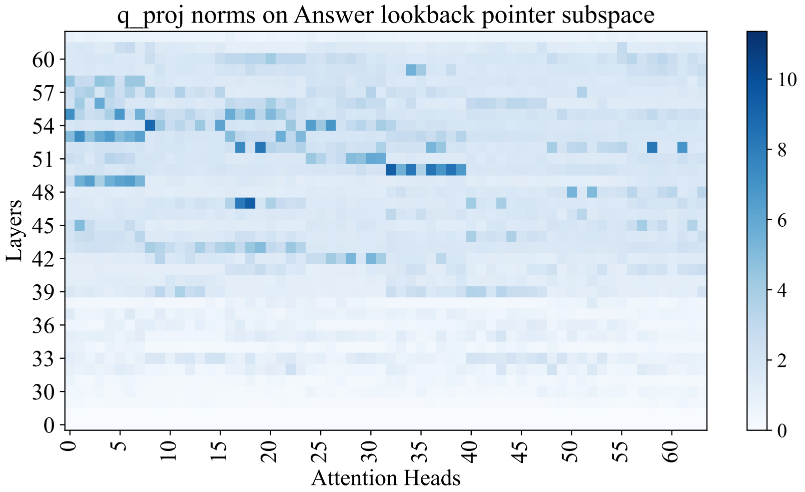


Figure 19: Alignment between the Answer lookback pointer causal subspace and the query projection matrix in Llama-3-70B-Instruct.

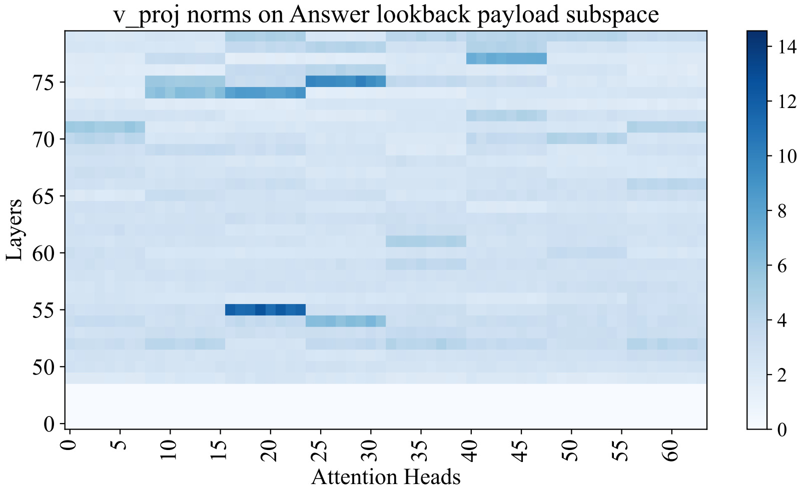
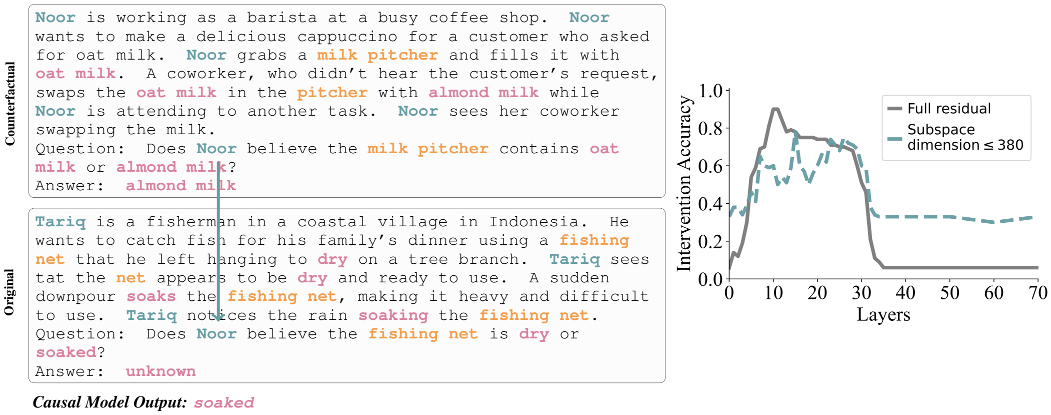


Figure 20: Alignment between the Answer lookback payload causal subspace and the value projection matrix in Llama-3-70B-Instruct.

queried character token in the original sample with the corresponding vector from the counterfactual sample (which contains the character OI), the model's output changes from unknown to the state token(s) associated with the queried object. This is because inserting the character OI at the queried token provides the correct pointer information, aligning with the address information at the correct state token(s), thereby enabling the model to form the appropriate QK-circuit and retrieve the state's OI. As shown in Fig. 21, we observe a high IIA between layers $9-28$—similar to the pattern seen in CausalToM—suggesting that the queried character token encodes the character OI in its residual vector within these layers.

Next, we investigate the Answer lookback mechanism in BigToM, focusing specifically on localizing the pointer and payload information at the final token position. To localize the pointer information, which encodes the correct state OI, we construct original and counterfactual samples by selecting two completely different examples from the BigToM dataset, each with different ordered states as the correct answer. For example, as illustrated in Fig.22, the counterfactual sample designates the first state as the answer, **thrilling plot**, whereas the original sample designates the second state, **almond milk**. We perform an intervention by swapping the residual vector at the last token position from the counterfactual sample into the original run. The causal model outcome of this intervention is that the



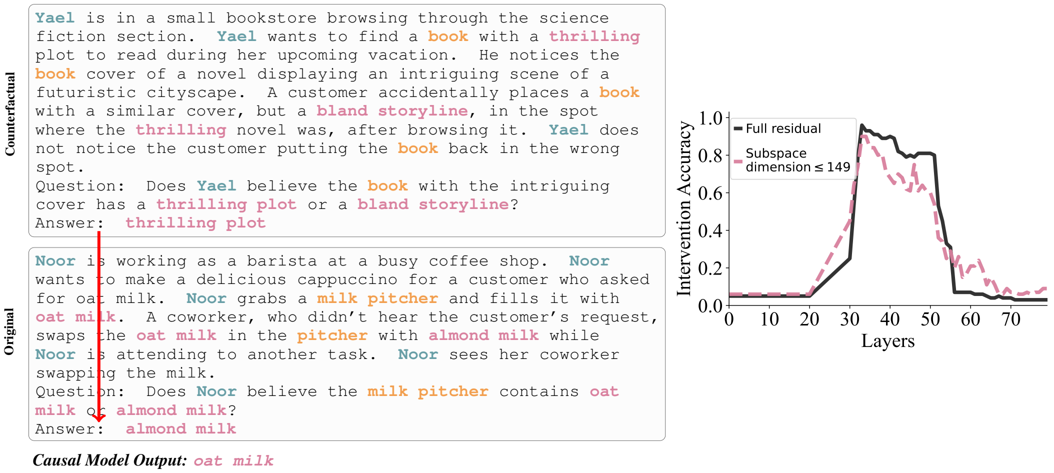
Noor is working as a barista at a busy coffee shop. Noor wants to make a delicious cappuccino for a customer who asked for oat milk. Noor grabs a milk pitcher and fills it with oat milk. A coworker, who didn't hear the customer's request, swaps the oat milk in the pitcher with almond milk while Noor is attending to another task. Noor sees her coworker swapping the milk.
Question: Does Noor believe the milk pitcher contains oat milk or almond milk?
Answer: almond milk

Tariq is a fisherman in a coastal village in Indonesia. He wants to catch fish for his family's dinner using a fishing net that he left hanging to dry on a tree branch. Tariq sees tat the net appears to be dry and ready to use. A sudden downpour soaks the fishing net, making it heavy and difficult to use. Tariq notices the rain soaking the fishing net.
Question: Does Noor believe the fishing net is dry or soaked?
Answer: unknown

Causal Model Output: soaked

Figure 21: **Query Character OI in BigToM**: This interchange intervention experiment inserts the first character's OI into the residual stream at the queried character token (●), resulting in the movement of pointer information to the last token that aligns with the address information of binding lookback mechanism. Consequently, the model is able to form the appropriate QK-circuit from the last token to predict the correct state answer token(s) as the final output, instead of unknown.

model will output the alternative state token from the original sample, **oat milk**. As shown in Fig.22, this alignment occurs between layers 33 and 51, similar to the layer range observed for the pointer information in the Answer lookback of CausalToM.



Yael is in a small bookstore browsing through the science fiction section. Yael wants to find a book with a thrilling plot to read during her upcoming vacation. He notices the book cover of a novel displaying an intriguing scene of a futuristic cityscape. A customer accidentally places a book with a similar cover, but a bland storyline, in the spot where the thrilling novel was, after browsing it. Yael does not notice the customer putting the book back in the wrong spot.
Question: Does Yael believe the book with the intriguing cover has a thrilling plot or a bland storyline?
Answer: thrilling plot

Noor is working as a barista at a busy coffee shop. Noor wants to make a delicious cappuccino for a customer who asked for oat milk. Noor grabs a milk pitcher and fills it with oat milk. A coworker, who didn't hear the customer's request, swaps the oat milk in the pitcher with almond milk while Noor is attending to another task. Noor sees her coworker swapping the milk.
Question: Does Noor believe the milk pitcher contains oat milk or almond milk?
Answer: almond milk

Causal Model Output: oat milk

Figure 22: **Answer Lookback Pointer in BigToM**: This interchange intervention experiment modifies the pointer information (●) of the Answer lookback, thereby altering the subsequent QK-circuit to attend to the other state (e.g., **oat milk**) instead of the original one (e.g., **almond milk**). As a result, the model retrieves the token value corresponding to the other state to answer the question.

Further, to localize the payload of the Answer lookback in BigToM, we perform an interchange intervention experiment using the same original and counterfactual samples as mentioned in the previous experiment, but with a different expected output—namely, the correct state from the counterfactual sample instead of the other state from the original sample. As shown in Fig. 23, alignment emerges after layer 59, consistent with the layer range observed for the Answer lookback payload in CausalToM.

Finally, we investigate the impact of the visibility condition on the underlying mechanism and find that, similar to CausalToM, the model uses the Visibility lookback to enhance the observing character's awareness based on the observed character's actions. To localize the effect of the visibility