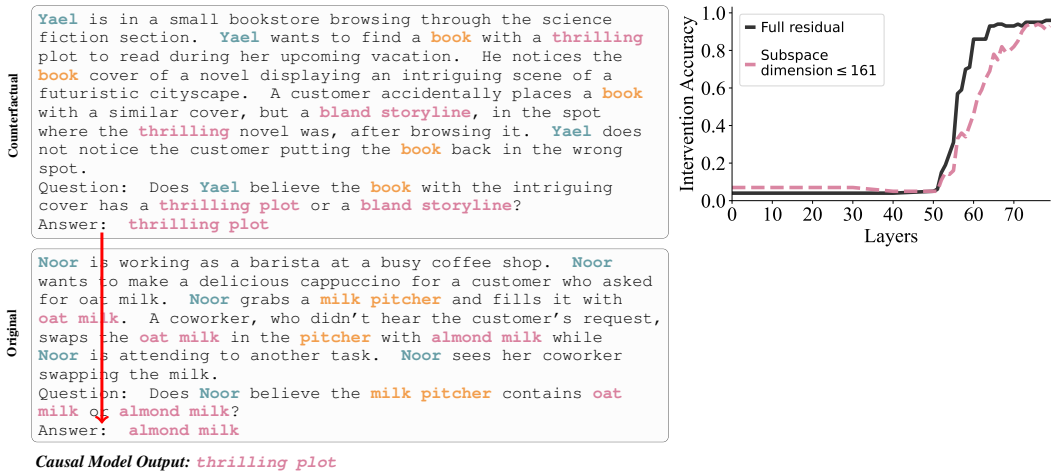

Figure 23: **Answer Lookback Payload in BigToM**: This interchange intervention experiment directly modifies the payload information (▲) of the Answer lookback, which is fetched from the corresponding state tokens and predicted as the next token(s). Thus, replacing its value in the original run, e.g. almond milk, with that from the counterfactual run, e.g. thrilling plot, causes the model's next predicted tokens to correspond to the correct answer of the counterfactual sample.

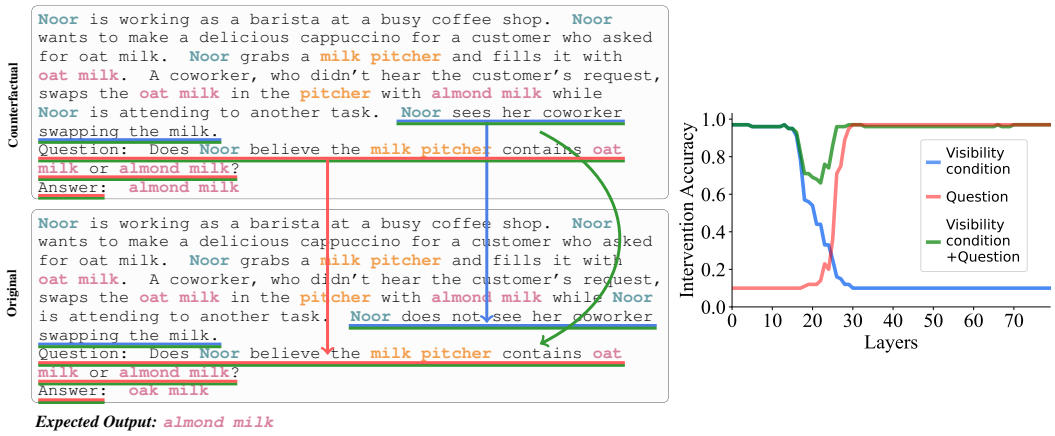

Figure 24: **Visibility Lookback in BigToM**: We perform three interchange interventions to establish the presence of the Visibility ID, which serves as both address and pointer information. When intervening at the source (●)—i.e., the visibility sentence—both the address and pointer are updated, resulting in alignment across layers. Intervening only at the subsequent question tokens leads to alignment only at later layers, after the model has already fetched the payload (▲). However, intervening at both the visibility and question sentences results in alignment across all layers, as the address and pointer remain consistent throughout.

condition, we perform an interchange intervention in which the original and counterfactual samples differ in belief type—that is, if the original sample involves a false belief, the counterfactual involves a true belief, and vice versa. The expected output of this experiment is the other (incorrect) state of the original sample. Following the methodology in Section 6, we conduct three types of interventions: (1) only at the visibility condition sentence, (2) only at the subsequent question sentence, and (3) at both the visibility condition and the question sentence. As shown in Fig. 24, intervening only at the visibility sentence results in alignment at early layers, up to layer 17, while intervening only at the subsequent question sentence leads to alignment after layer 26. Intervening on both the visibility and question sentences results in alignment across all layers. These results align with those found in the CausalToM setting shown in the Fig. 8.

Previous experiments suggest that the underlying mechanisms responsible for answering belief questions in BigToM are similar to those in CausalToM. However, we observed that the subspaces encoding various types of information are not shared between the two settings. For example, although the pointer information in the Answer lookback encodes the correct state's OI in both cases, the specific subspaces that represent this information at the final token position differ significantly. We leave a deeper investigation of this phenomenon—shared semantics across distinct subspaces in different distributions—for future work.

L Generalization of Belief Tracking Mechanism on CausalToM to Llama-3.1-405B-Instruct

This section presents all the interchange intervention experiments described in the main text, conducted using the same set of counterfactual examples on Llama-3.1-405B-Instruct, using NDIF Fiotto-Kaufman et al. (2025). Each experiment was performed on 80 samples. Due to computational constraints, subspace interchange intervention experiments were not conducted. The results indicate that Llama-3.1-405B-Instruct employs the same underlying mechanism as Llama-3-70B-Instruct to reason about belief and answer related questions. This suggests that the identified belief-tracking mechanism generalizes to other models capable of reliably performing the task.

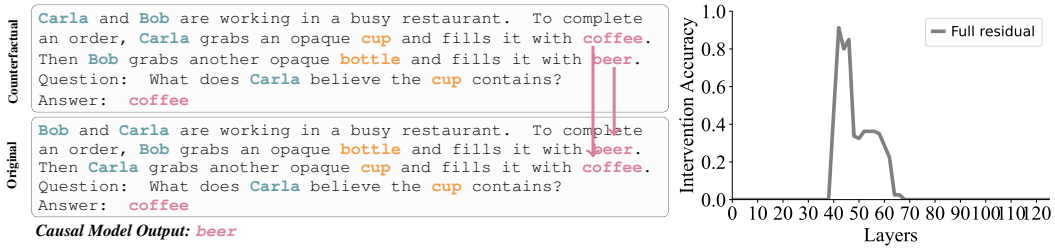

Figure 25: **Payload and address of Binding lookback**

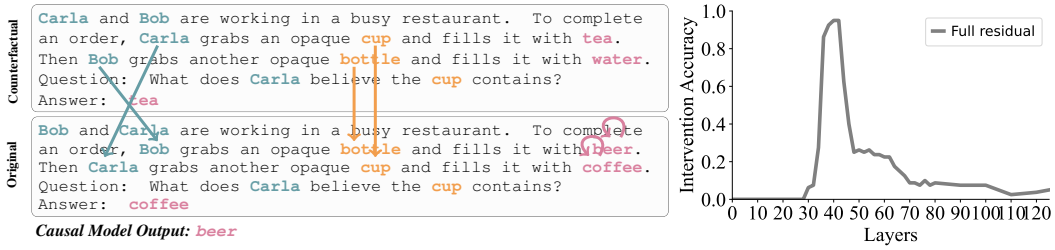

Figure 26: **Source Information of Binding lookback**

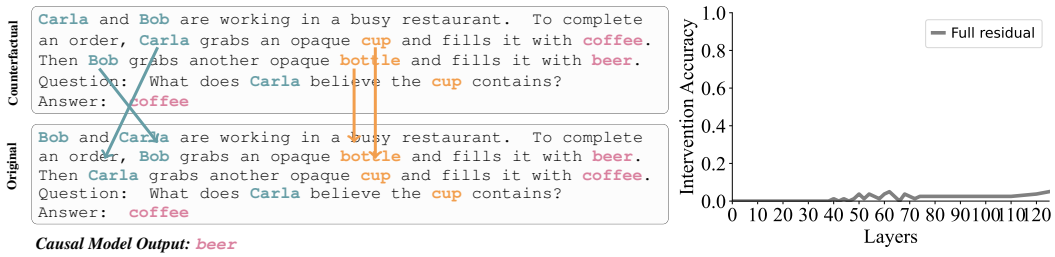

Figure 27: **Source Reference Information of Binding lookback without freezing address and payload**

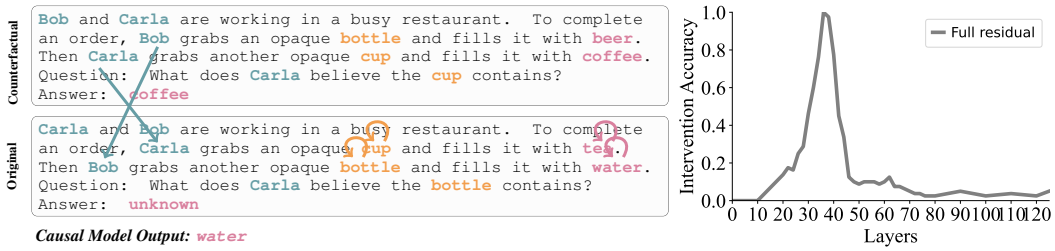

Figure 28: **Character OI**

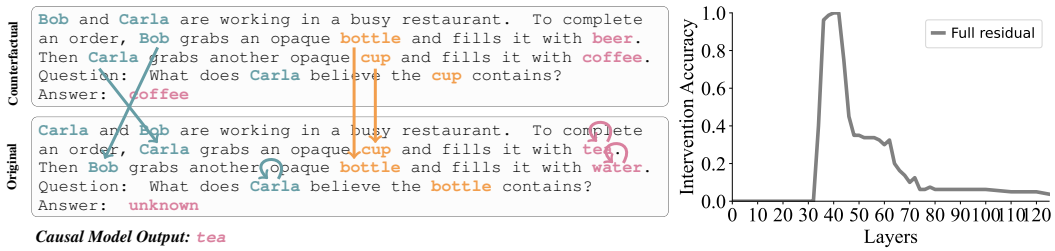

Figure 29: **Object OI**

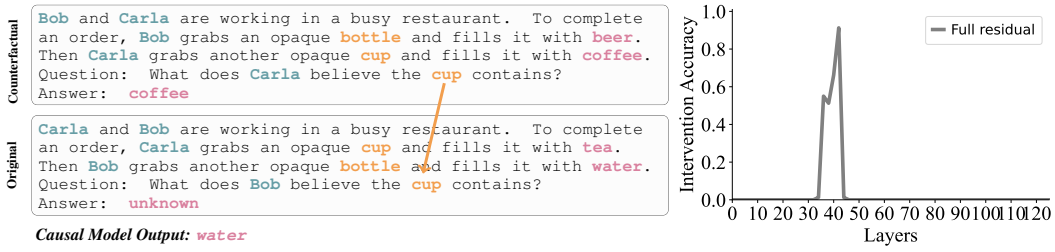

Figure 30: **Query Object OI**

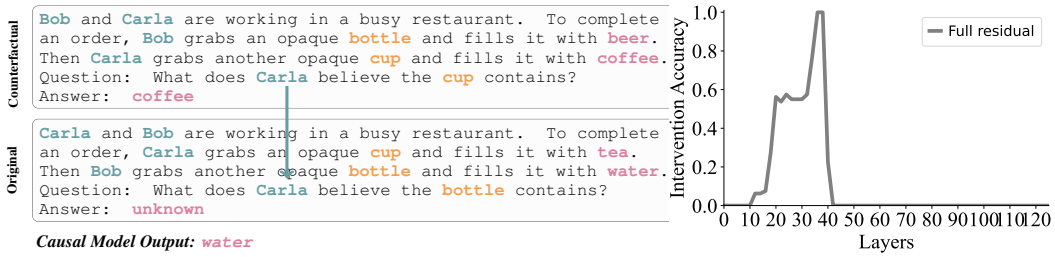

Figure 31: **Query Character OI**