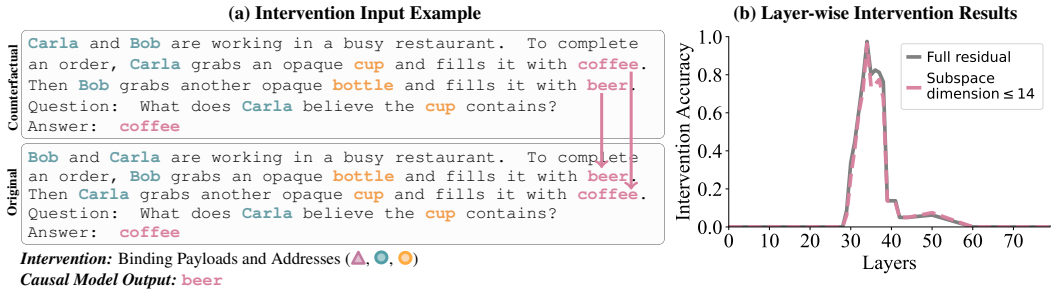

Figure 5: **Binding lookback Address and Payload:** The causal model predicts that swapping addresses (character and object OIs; and) and payloads (state OIs;) should cause the binding lookback mechanism to attend to the alternate state token and retrieve its state OI. This retrieved state OI is then dereferenced by the answer lookback, producing the corresponding token as the output (e.g., **beer** instead of **coffee**). The LM's behavior matches this prediction when we perform interchange interventions on the state token across layers 33–38. These findings support our hypothesis that both address and payload information are encoded in the residual stream of state tokens.

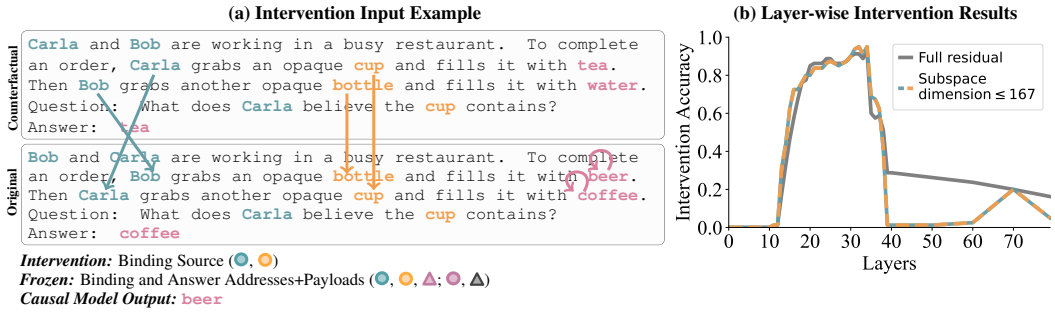

Figure 6: **Source Reference Information of Binding lookback**: The causal model predicts that swapping the source reference information (character and object OIs; ,), while freezing the addresses and payloads of the binding lookback, should cause the binding lookback mechanism to attend to the alternate state token and retrieve its state OI, which would generate alternate state token as the final output via the answer lookback (e.g., **beer** instead of **coffee**). The LM's behavior matches this prediction when we perform interchange interventions at the character and object tokens across layers 20-34. These results support our hypothesis that source reference information is encoded in the residual stream of character and object tokens.

In the LM, we interchange the residual streams of the character and object tokens layer-by-layer, while keeping the residual stream of the state token fixed. As shown in Fig. 6b, this experiment reveals alignment between layers 20 and 34, indicating that source reference is encoded in the residual streams of the character and object tokens within this layer range. Additional results are provided in Appendix G, where Fig. 13 shows that freezing the residual stream of the state token is necessary for this alignment to emerge. These findings support our hypothesis that source reference is present in the character and object tokens and is subsequently transferred to the recalled and lookback tokens.

Localizing the Pointer Information Finally, we localize the pointer copies of the character and object OIs to their corresponding tokens in the question and to the final token. See Appendices G & H for details of the experiments and results.

In summary, belief tracking begins in layers 20–34, where character and object OIs are encoded in their respective token representations. These OIs are transferred to the corresponding state tokens in layers 33–38. When a question is asked, pointer copies of the relevant character and object OIs are moved to the final token by layer 34, where they are dereferenced to retrieve the correct state OI. At the final token, this state OI is represented across layers 34–52, and between layers 52–56, it is dereferenced to fetch the answer from the correct state token, producing the final output.

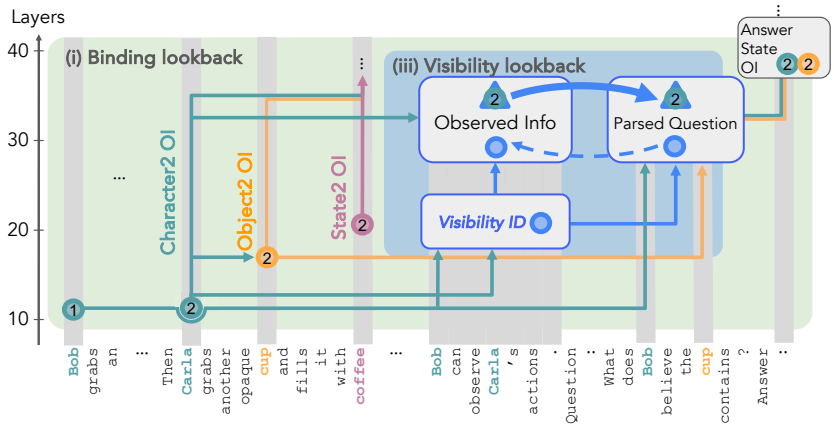

Figure 7: **Visibility Lookback** When one (observing) character can see another (observed) character, the LM assigns a visibility ID (●) to the visibility sentence where this relation is defined. An address copy of this visibility ID remains in the visibility sentence's residual stream. A pointer copy of the visibility ID is transferred to the subsequent tokens' residual stream. The LM dereferences this pointer through a QK-circuit, bringing forward the payload (▲), when processing subsequent tokens. Based on initial evidence, this payload contains the observed character's OI(●). See Appendix I for details. This mechanism allows the model to incorporate the observed character's knowledge into the observing character's belief state, enabling more complex belief reasoning.

6 Impact of Visibility Conditions on Belief Tracking Mechanism

So far, we have demonstrated how the LM uses ordering IDs and two lookback mechanisms to track the beliefs of characters that cannot observe each other. Now, we explore how the LM updates the beliefs of characters when one character (*observing*) can observe the actions of the other (*observed*).

Hypothesized Visibility Lookback Mechanism We hypothesize that the LM uses an additional lookback mechanism, which we call the *Visibility Lookback*, to integrate information about the observed character when it is explicitly stated that one character can see another's action. As illustrated in Fig. 7, we hypothesize that the LM first generates a *Visibility ID* (●) at the residual stream of the visibility sentence, serving as the source reference information. The address copy of the visibility ID remains in the residual stream of the visibility sentence, while its pointer copy gets transferred to the residual stream of the subsequent tokens (lookback tokens). Then LM forms a QK-circuit at the lookback tokens and dereferences the visibility ID pointer to retrieve the payload.

Although our two-character setting is unable to discern the exact semantics of the payload in the visibility lookback, our observations are consistent with the payload encoding the observed character's OI. Our initial observations suggest another lookback where the story sentence associated with the observed character serves as the source reference, and its payload encodes information about the observed character. The observed character's OI appears to be retrieved by the lookback tokens of the Visibility lookback, with causal effects on the queried character's awareness (see App. I for details).

6.1 Verifying the Hypothesized Visibility Lookback

Localizing the Source Reference In this experiment, we localize the Visibility ID (●), i.e., the source reference of the Visibility lookback. We conduct an interchange intervention experiment where the counterfactual is a different story in which the characters' visibility is flipped from unobserved to observed (Fig. 8a), and we look for an output change from "unknown" to the answer that would be observed. We intervene on the representation of all the visibility sentence tokens. As shown in Fig. 8b (blue — line), causal effects appear between layers 10 and 23, indicating that the visibility ID remains encoded in the visibility sentence until layer 23, after which it is split into address and pointer copies that must be connected by dereference to have an effect. This pattern supports our hypothesis that the LM generates a reference to the Visibility ID.

Localizing the Payload Next, we localize the payload (▲) information using the same counterfactual dataset. However, instead of intervening on the recalled tokens, we intervene on the lookback

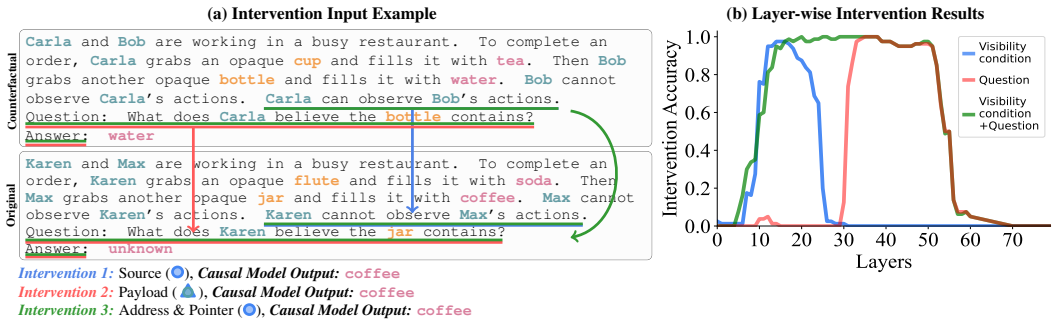

Figure 8: **Visibility Lookback**: We conduct three interchange intervention experiments to support the Visibility Lookback hypothesis: (1) *Source Alignment*: We align the source reference ( ) by intervening on the visibility sentence, replacing it with its representation from a counterfactual run where the visibility sentence causes the queried character to become aware of the queried object's contents. We observe that source reference information aligns between layers 10 and 23, after which it splits into separate address and pointer components. (2) *Payload Alignment*: To align the payload ( ), we intervene on all subsequent tokens and observe alignment only after layer 31. (3) *Address and Pointer Alignment*: When intervening on both the address and pointer information ( ), we observe alignment across a broader range of layers, particularly between layers 24 and 31, because of the enhanced alignment between the address and pointer copies at the recalled and lookback tokens.

tokens, specifically the question and answer tokens. As in the previous experiment, we replace the residual vectors of these tokens in the original run with those from the counterfactual run. As shown in Fig. 8b (red — line), alignment occurs after layer 31, indicating that the information causing the queried character's awareness is present in the lookback tokens after this layer.

Localizing Address and Pointer The previous two experiments indicate the absence of both the source and payload information between layers 24 and 31. We hypothesize that this lack of signal is due to a mismatch between the address and pointer information that inhibits a lookback dereference. Specifically, when intervening only on the recalled token after layer 25, the pointer is not updated, whereas intervening only on the lookback tokens leaves the address unaltered, a mismatch in either case. To test this hypothesis, we conduct another intervention using the same counterfactual dataset, but this time, we intervene on the residual vectors of both the recalled and lookback tokens, i.e., the visibility sentence, as well as the question and answer tokens. As shown in Fig. 8b (green — line), alignment occurs after layer 10 and remains stable, providing evidence that a lookback now occurs between layers 24 and 31. This intervention replaces both the address and pointer copies of the visibility IDs, enabling the LM to form a QK-circuit and resolve the visibility lookback.

7 Related Work

Theory of mind in LMs Theory of mind in LMs has been widely benchmarked (Le et al., 2019; Shapira et al., 2023; Wu et al., 2023; Kim et al., 2023; Xu et al., 2024; Jin et al., 2024; Chan et al., 2024; Strachan et al., 2024b). However, these benchmarks lack adequate counterfactuals for the binding manipulations we need, so we made CausalToM (Section B).

Entity tracking in LMs Entity tracking and variable binding are crucial abilities for LMs to exhibit not only coherent ToM capabilities, but also neurosymbolic reasoning. Many existing works have attempted to decipher this ability in LMs (Li et al., 2021; Davies et al., 2023; Feng & Steinhardt, 2023; Kim & Schuster, 2023; Prakash et al., 2024; Feng et al., 2024; Dai et al., 2024; Wu et al., 2025). Our work builds on their empirical insights and extends the current understanding of how LMs bind various entities defined in context.

Mechanistic interpretability of theory of mind Few studies explored the underlying mechanisms of ToM of LM (Zhu et al., 2024; Bortoletto et al., 2024; Herrmann & Levinstein, 2024). Those studies use probing (Alain, 2016; Belinkov, 2022) to identify internal representations of beliefs and steering (Rimsky et al., 2023; Li et al., 2024) to control LMs by manipulating their activations. However, the mechanism by which LMs solve those tasks remains a black box, limiting our ability to understand, predict, and control LMs' behaviors.