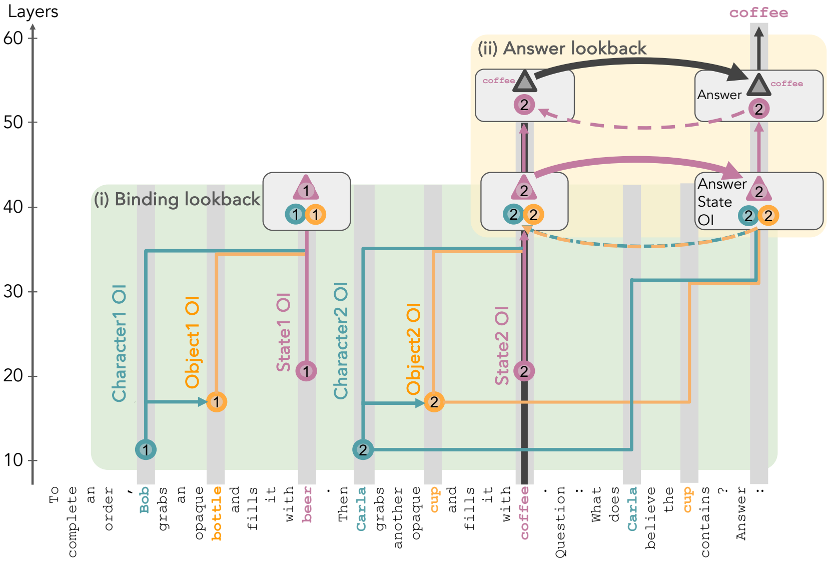

Layers
60
50
40
30
20
10
coffee
(ii) Answer lookback
coffee
2
Answer
coffee
2
2
Answer
State
OI
2
2 2
2 2
(i) Binding lookback
1
1 1
Character1 OI
Object1 OI
State1 OI
Character2 OI
Object2 OI
State2 OI
1
1
1
2
2
2
To complete an order , Bob grabs an opaque bottle and fills it with beer . Then Carla grabs another opaque cup and fills it with coffee . Question : What does Carla believe the cup contains ? Answer :

Figure 3: **Belief Tracking with *no visibility* between characters.** We hypothesize that the LM tracks beliefs using two lookback mechanisms. First, in (i) **Binding lookback**, LM binds together each character-object-state triple in the state token residual stream. When asked about a specific character-object pair, the LM looks back to the corresponding OIs to retrieve the correct state OI. Second, in (ii) **Answer lookback**, LM dereferences that state OI (used as a pointer) to retrieve the token value of the correct state. Colors indicate information type, shapes indicate role of information in lookback (see Fig. 1), e.g., state OI is a payload (▲) in (i) and a pointer-address (●) in (ii).

2024; Vig et al., 2020; Meng et al., 2022), several significant insights emerge: 1) Information from the correct state token (**beer**) flows directly from its residual stream to that of the final token in later layers, consistent with prior findings (Lieberum et al., 2023; Prakash et al., 2024); 2) Information associated with the query character and the query object is retrieved from their earlier occurrences and passed to the final token before being replaced by the correct state token.

Desiderata Based Patching via Causal Abstraction The causal mediation experiments provide a coarse-grained analysis of where information flows, but do not identify what information is being transferred. In a transformer, the first layer represents the input and the last layer represents the output, but we wish to know: what is represented in the middle? We analyze the internal mechanism using *Causal Abstraction* (Geiger et al., 2021; 2024); First, we hypothesize a high-level causal model of the computational steps from input to output (Sec. 4), and then align its variables with the LM's internal activations (Sec. 5). We test the alignment through targeted interchange interventions on causal variables in the hypothesized model and hidden activations in the LM. If the LM produces the same output as the causal model under these aligned interventions, it provides evidence supporting the hypothesized causal model. We quantify this effect using *interchange intervention accuracy* (IIA; Geiger et al., 2022), which measures the proportion of cases where the intervened causal model and intervened LM agree. See Appendix D for more details.

In addition to measuring IIA on entire residual stream vectors, we also intervene on localized subspaces to further isolate causal variables. To identify the subspace of a specific variable, we employ *Desiderata-based Component Masking* (De Cao et al., 2020; Davies et al., 2023; Prakash et al., 2024). This method learns a sparse binary mask over the activation space that maximizes the logit of the hypothesized causal model output. We train a mask to select singular vectors of the activation space that encode a high-level variable (see Appendix F for details). Our experiments in Sec. 5 report both interventions on the full residual stream and on the identified subspaces.

4 Hypothesized High-Level Causal Model of Belief Tracking

Here we start with an overview of our hypothesized causal model of belief-tracking when characters are not aware of each other's actions. The causal model is an algorithmic process that has variables with structural roles that do not refer to the details of a transformer architecture. Appendix E presents the full pseudocode of the causal model. In Section 5, we will present experiments to verify that the causal model's variables align with representations in the transformer.

Belief tracking begins when the causal model assigns *ordering IDs* (OIs; ●, ●, ●) to each character, object, and state token, marking their order of appearance. For instance, in the example in Fig. 3, `Bob` is assigned first character OI (①), and `Carla` is assigned the second character OI(②). Then it uses these OIs in two lookback mechanisms:

(i) Binding lookback. The causal model creates address copies of each character OI (●) and object OI (●) that are bound to the state OI (Binding Payload, ▲), creating a character–object–state triple. When a question is asked about a character and object, the causal model creates pointer copies of that character and object OIs (Binding Pointers ●, ●) and dereferences them to retrieve the state OI.

(ii) Answer lookback. The causal model creates an address copy of the state OI (Answer Address ●) that is bound to the state token (Answer Payload, ▲). Through the binding lookback, a pointer copy of this OI (●) is created. The causal model dereferences the pointer to retrieves the correct state token payload as the final output.

5 Verifying the Hypothesized Causal Model of Belief Tracking

We test our hypothesized causal model by localizing its variables within the transformer's neural representations. Specifically, we localize the addresses, pointers, and payloads of the (i) binding lookback and (ii) answer lookbacks within the LM's internal activations. In Fig. 3, we show a trace of the causal model run on an input overlayed onto a schematic of a transformer architecture. This visualizes the alignment between variables in the causal model and locations in the LM residual stream that the experiments in the remaining of this paper will support. In the binding lookback, the character and object OI addresses are realized in the residual stream of the state token. The pointer copies are brought forward to the last token residual stream where they are dereferenced via attention to bring forward the correct state OI payload. In the answer lookback, the address copy of the state OI is in the state token residual stream while the pointer copy is in the last token residual stream.

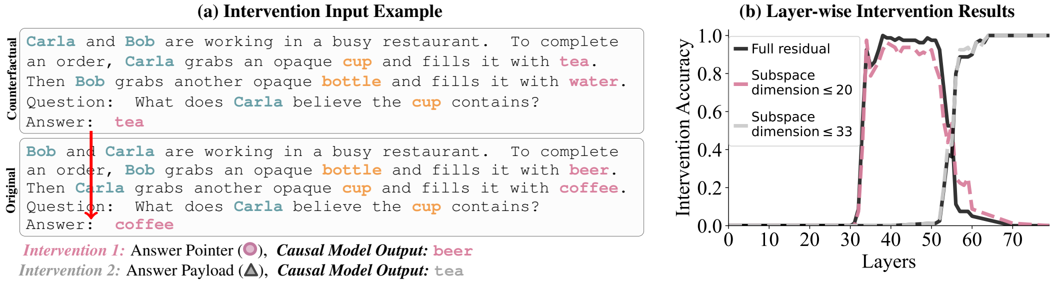
Each of the following experiments localizes the presence of specific ordering IDs (OIs) and verifies their roles as hypothesized by our causal model. We do this by targeted interchange intervention experiments on the causal model and the LM. We copy hidden states between identical tokens (for example, replacing the representation of "**:**" in one context with the representation of "**:**" in another context, as in Fig. 4). When this intervention causes the LM's have the same output as the causal model under an interchange intervention on OI variables, we have evidence that the OI is carrying out the hypothesized role. Each experiment reports the effects of $n = 80$ different cases with the same structure, and the effect is measured at every layer.

Because the last step of the causal model is easiest to understand, we proceed through the experiments in reverse order, beginning with an experiment to verify the final "answer lookback" stage. After this instructive starting point, we work backward to verify the earlier steps of the model. Additional results can be found in Appendix G and H.

5.1 Step II: Answer Lookback – Retrieving the Correct State

Localizing the Answer Payload We first verify the presence of the correct Answer Payload ▲ at the deepest layer representation of final token "**:**". To do so, we run an interchange intervention experiment shown in Fig. 4a in which the counterfactual example **c** swaps the order of the characters and objects of the original example **o** and also replaces the state (drinks) tokens with new values. If the Answer Payload is correctly localized, swapping it should cause the answer of the counterfactual (e.g., tea) to replace the answer of the original example (e.g., coffee). The gray line in Fig. 4b shows that this output change is observed in every one of $n = 80$ cases, both when intervening on the full residual stream and on the identified subspace. However, not at every layer: the information is only present after layer 56, indicating that before this stage, the transformer has not yet retrieved the correct answer payload into the residual stream. That is consistent with our hypothesis that at early steps, the OI has not yet been dereferenced. At an earlier stage, we expect to see an Answer Pointer.

Localizing the Pointer Information To identify the Answer Pointer ● before it is dereferenced to bring the payload (state token value), we examine the representations of "**:**" at layers earlier than 56.



(a) Intervention Input Example

Counterfactual

```
Carla and Bob are working in a busy restaurant.  To complete
an order, Carla grabs an opaque cup and fills it with tea.
Then Bob grabs another opaque bottle and fills it with water.
Question:  What does Carla believe the cup contains?
Answer:  tea
```

Original

```
Bob and Carla are working in a busy restaurant.  To complete
an order, Bob grabs an opaque bottle and fills it with beer.
Then Carla grabs another opaque cup and fills it with coffee.
Question:  What does Carla believe the cup contains?
Answer:  coffee
```

Intervention 1: Answer Pointer (●), ***Causal Model Output:*** beer

Intervention 2: Answer Payload (▲), ***Causal Model Output:*** tea

(b) Layer-wise Intervention Results

Figure 4: **Answer Lookback Pointer and Payload**: The causal model predicts that if we alter the "Answer Payload ▲" of the original to instead take the value of the counterfactual answer payload, the output should change from coffee to tea; the gray curve in the line plot shows this does occur when patching residual vectors at the ":" token beyond layer 56, providing evidence that the answer payload resides in those states. On the other hand the causal model predicts that taking the counterfactual "Answer Pointer ●" would change the original run output from coffee to beer—a new output that matches *neither* the original nor the counterfactual!—and we do see this surprising effect, again when patching layers between 34 and 52, providing strong evidence that the answer pointer is encoded at those layers. These results suggest the Answer Lookback occurs between layers 52 and 56.

Our causal model provides the hypothesis: if the Answer Pointer is present, then patching the pointer from the counterfactual run into the original run should redirect the LM to attend to the location of the correct counterfactual state and fetch its payload. For example, in Fig. 4a the counterfactual pointer references the first presented state. When we patch it into the original story, we expect the model's answer to change to beer rather than coffee. The colored line in Fig. 4b confirms that this effect is consistently observed when patching any layer between $34 - 52$ (both when patching the full residual stream and the identified subspace), supporting our hypothesis that these layers encode the Answer Pointer information at the final token, rather than directly transferring token values.

5.2 Step I: Binding Lookback – Linking Characters, Objects, and States

Localizing the Address and Payload In this experiment, we verify the presence of the address copies of the character and object OIs as well as the payload (state OI) at the state token residual stream (recalled token, Fig. 3). As illustrated in Fig. 5a, we construct an intervention dataset where each example consists of an original input $\mathbf{o}$ with an answer that is not *unknown* and a counterfactual input $\mathbf{c}$ where the character, object, and state token values are identical, except the ordering of the two story sentences is swapped while the question remains unchanged. The expected LM's output predicted by our hypothesized causal model is the other state token in the original example, e.g., beer. That is because patching the address and payload values at each state token, without changing the pointer, makes the LM dereference the other state token. As a result, the model's output should flip to the other state token in the original input.

We perform the interchange intervention experiment layer-by-layer, where we replace the residual stream vector (or the identified subspace) of the first state token in the original run with that of the second state token in the counterfactual run and vice versa for the other state token. It is important to note that if the intervention targets state token values instead of their OIs, it should not produce the expected output. (This happens in the earlier layers.) As shown in Fig. 5b, the strongest alignment occurs between layers 33 and 38, supporting our hypothesis that the state token's residual stream contains both the address (character and object OIs) and the payload information (state OI).

Localizing the Source Reference Information Next, we localize the source reference information, i.e., character and object OIs at their respective token residual stream. As illustrated in Fig. 6a, we conduct an intervention experiment with a dataset where the counterfactual example, $\mathbf{c}$, swaps the order of the characters and objects as well as replaces the state tokens with entirely new ones while keeping the question the same as in $\mathbf{o}$. Under this setup, an interchange intervention on the hypothesized causal model that targets the source reference should propagate changes through both the address and the pointer, leaving the final output unchanged. However, if we instead freeze the state token residual stream, which carries both the payload and the address, the causal model produces the alternate state token (e.g., beer in Fig. 6), as the pointer refers to the other state's address.