
Language Models use Lookbacks to Track Beliefs

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 How do language models (LMs) represent characters’ beliefs, especially when those
2 beliefs may differ from reality? This question lies at the heart of understanding the
3 Theory of Mind (ToM) capabilities of LMs. We analyze Llama-3-70B-Instruct’s
4 ability to reason about characters’ beliefs using causal mediation and abstraction.
5 We construct a dataset that consists of simple stories where two characters each
6 separately change the state of two objects, potentially unaware of each other’s
7 actions. Our investigation uncovered a pervasive algorithmic pattern that we call a
8 *lookback mechanism*, which enables the LM to recall important information when
9 it becomes necessary. The LM binds each character-object-state triple together by
10 co-locating reference information about them, represented as their Ordering IDs
11 (OIs) in low rank subspaces of the state token’s residual stream. When asked about
12 a character’s beliefs regarding the state of an object, the *binding lookback* retrieves
13 the corresponding state OI and then an *answer lookback* retrieves the state token.
14 When we introduce text specifying that one character is (not) visible to the other,
15 we find that the LM first generates a *visibility ID* encoding the relation between
16 the observing and the observed character OIs. In a *visibility lookback*, this ID is
17 used to retrieve information about the observed character and update the observing
18 character’s beliefs. Our work provides insights into the LM’s belief tracking
19 mechanisms, taking a step toward reverse-engineering their ToM capabilities.

20 1 Introduction

21 The ability to infer the mental states of others—known as Theory of Mind (ToM)—is an essential
22 aspect of social and collective intelligence [Premack and Woodruff, 1978, Riedl et al., 2021]. Recent
23 studies have established that language models (LMs) can solve some tasks requiring ToM reasoning
24 [Street et al., 2024, Strachan et al., 2024b, Kosinski, 2024], while others have highlighted shortcomings
25 [Sclar et al., 2025, Shapira et al., 2024, Kim et al., 2023a, *inter alia*]. Nonetheless, most existing
26 work relies on behavioral evaluations, which do not shed light on the internal mechanisms by which
27 LMs encode and manipulate representations of mental states to solve (or fail to solve) such tasks [Hu
28 et al., 2025, Gweon et al., 2023].

29 In this work, we investigate how LMs represent and update characters’ beliefs, which is a fundamental
30 element of ToM [Dennett, 1981, Wimmer and Perner, 1983]. For instance, the Sally-Anne test [Baron-
31 Cohen et al., 1985], a canonical test of ToM in humans, evaluates this ability by asking individuals to
32 track Sally’s belief, which diverges from reality due to missing information, and Anne’s belief, which
33 updates based on new observations.

34 We construct *CausalToM*, a dataset of simple stories involving two characters, each interacting with
35 an object to change its state, with the possibility of observing one another. We then analyze the
36 internal mechanisms that enable Llama-3-70B-Instruct [Grattafiori et al., 2024] to reason about and
37 answer questions regarding the characters’ beliefs about the state of each object (for a sample story,
38 see Section 3 and for the full prompt refer to Appendix A).

We discover a pervasive computation that performs multiple subtasks, which we refer to as the *lookback mechanism*. This mechanism enables the model to recall important information only when it becomes necessary. In a lookback, two copies of a single piece of information are transferred to two distinct tokens. This allows attention heads at the latter token to look back at the earlier one when needed and retrieve vital information stored there, rather than transferring that information directly.

We identify three key lookback mechanisms that collectively perform belief tracking: 1) *Binding lookback* (Fig. 2a(i)): First the LM assigns *ordering IDs* (OIs) [Dai et al., 2024] that encode whether a character, object, or state token appears first or second. Then, the character and object OIs are copied to low-rank subspaces of the corresponding state token and the final token residual stream. Later, when the LM needs to answer a question about a character’s beliefs, it uses this information to retrieve the answer state OI. 2) *Answer lookback* (Fig. 2a(ii)): Uses the answer state OI from the binding lookback to retrieve the answer state token value. 3) *Visibility lookback* (Fig. 6): When an explicit visibility condition between characters is mentioned, the model employs additional reference information called the *visibility ID* to retrieve information about the observed character, augmenting the observing character’s awareness.

Overall, this work not only advances our understanding of the internal computations in LMs that enable ToM capability but also uncovers a pervasive mechanism that serves as the foundation for executing complex logical reasoning with conditionals.

2 The Lookback Mechanism

Our investigation of belief tracking uncovers a recurring pattern of computation that we call the *lookback mechanism*.¹ Here we give a brief overview of this mechanism; subsequent sections provide detailed experiments and analyses. In lookback, *source information* is copied (via attention) into an *address* copy in the residual stream of a *recalled token* and a *pointer* copy in the residual stream of a *lookback token* that occurs later in the text. The LM places the address alongside a *payload* of the recalled token’s residual stream that can be brought forward to the lookback token if necessary. Fig. 1 schematically describes a generic lookback.

That is, the LM can use attention to dereference the pointer and retrieve the payload present in the residual stream of the recalled token (that might contain aggregated information from previous tokens), bringing it to the residual stream of the lookback token. Specifically, the pointer at the lookback token forms an attention query vector, while the address at the recalled token forms a key vector. Because the pointer and the address are copies of the same source information, they would have a high dot-product, hence a *QK-circuit* [Elhage et al., 2021] is established forming a bridge from the lookback token to the recalled token. The LM uses this bridge to move the payload that contains information needed to complete the subtask through the *OV-circuit*.

To develop an intuition for why an LM would learn to implement lookback mechanisms to solve reasoning tasks such as our belief tracking task, consider that during training LMs process text in sequence with no foreknowledge of what might come next. Then, it would be useful to mark

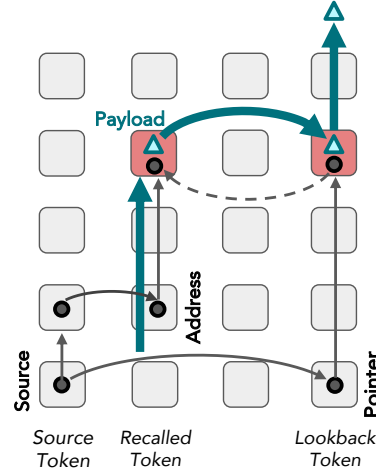


Figure 1: **The lookback mechanism** is used to perform conditional reasoning. The *source token* contains information that is copied into two instances via attention to create a *pointer* and an *address*. Alongside the address in the residual stream is a *payload* of information. When necessary, the model retrieves the payload by dereferencing the pointer. Solid lines represent movement of information, while the dotted line indicates the attention “looking back” from pointer to address.

¹Although this mechanism may resemble *induction heads* [Elhage et al., 2021, Olsson et al., 2022], they differ fundamentally. In induction heads, information from a previous token occurrence is passed only to the subsequent token through, without being duplicated to its next occurrence. In contrast, the lookback mechanism copies the same information not only to the location where the vital information resides but also to the target location that needs to retrieve that information.

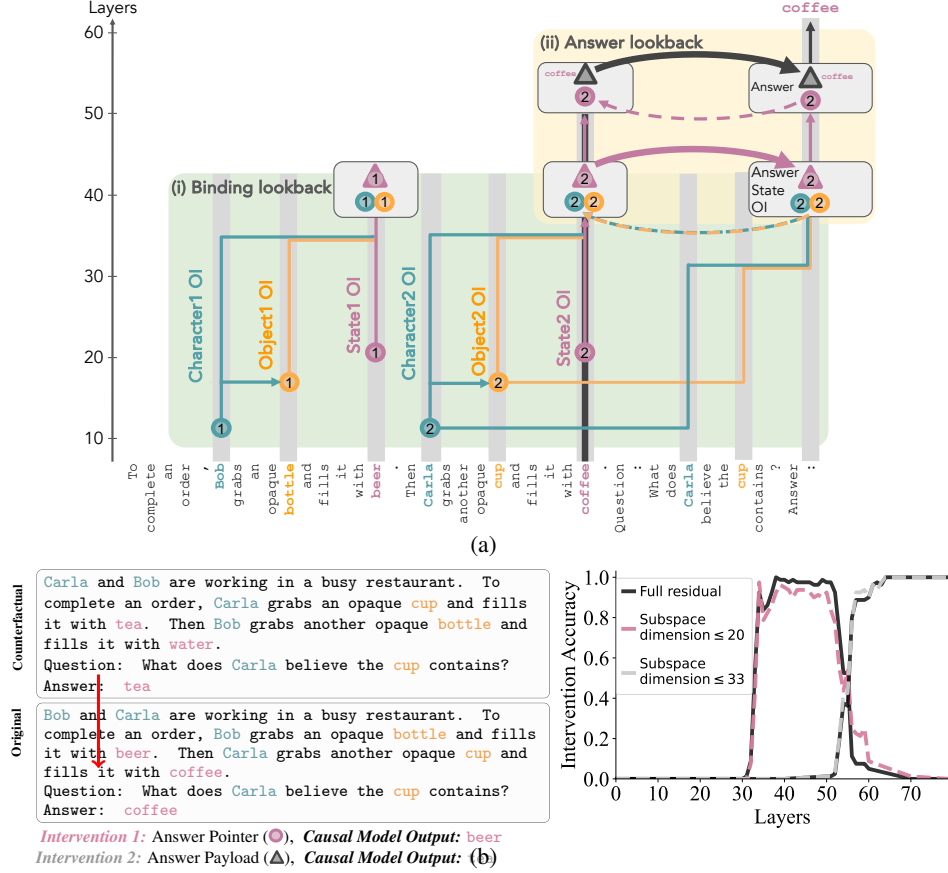


Figure 2: **Belief Tracking in Language Models:** We task the LM with tracking the beliefs of two characters that manipulate the states of two objects. We hypothesize that the LM solves this task by implementing a causal model with two lookback mechanisms (2a). To support our hypothesis, we conduct a causal analysis where we measure whether interventions on a high-level causal model produce the same output as equivalent interventions on the LM. For instance in 2b, we show results for an experiment distinguishing the pointer and payload in the answer lookback.

(a) **Belief Tracking with No Visibility between Characters:** Our hypothesized causal model for this kind of story has two lookbacks that operate on ordering IDs (OIs) that encode whether a token appears first or second. In the **binding lookback** (i), the LM first represents the two events in the story by binding together each character-object-state triple in the residual stream of the state token. When questioned about a particular character and object, the LM looks back to the corresponding triple and retrieves an OI to that state token. Notice that in this lookback, that payload is later used as a pointer, i.e., what a C programmer would call a double pointer. In the **answer lookback** (ii), the LM dereferences the pointer to the answer token to generate the correct answer. Color indicates the information content, while shape indicates the role of that information in lookback (see Fig. 1), e.g., the state OI is a payload (\blacktriangle) in the binding lookback and a pointer/address (\odot) in the answer lookback.

(b) **Answer Lookback Pointer and Payload:** To test our hypothesized causal model, we run the LM on pairs of slightly different stories, and then intervene by patching a specific representation state from the counterfactual run to the original run, observing any change in the output. The causal model predicts that if we alter the “answer payload \blacktriangle ” 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 with $p \approx 1.0$ when patching states 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 \odot ” 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 with $p \approx 1.0$, when patching layers between 34 and 52, providing strong evidence that the answer pointer is encoded at those layers. Collected over $N = 80$ samples, these measurements suggest the Answer Lookback occurs between layers 52 and 56. Furthermore the representations of the causal variables are small: the interventions can be localized even further to subspaces of dimension 33 (payload) or 20 (pointer), tiny portions of the 8192-dimensional state space.

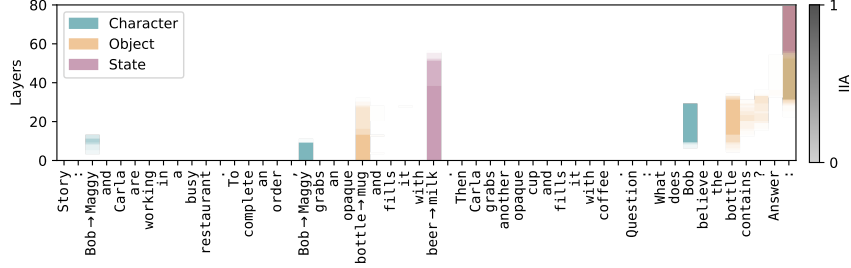


Figure 3: Tracing information flow of crucial input tokens using causal mediation analysis.

addresses alongside payloads that might be useful for downstream tasks. In our setting, the LM constructs a representation of the story without any knowledge of what questions it may be asked about, so the LM concentrates pieces of information in the residual stream of certain tokens which later become payloads and addresses. When the question text is reached, pointers are constructed that reference this crucial story information and dereference it as the answer to the question.

3 Preliminaries

Dataset Existing datasets for evaluating ToM capabilities of LMs are designed for behavioral testing and lack counterfactual pairs needed for causal analysis [Kim and Sundar, 2012]. To address this, we constructed *CausalToM*, a structured dataset of simple stories, where each story involves two characters, each interacting with a distinct object causing the object to take a unique state. For example: “**Character1** and **Character2** are working in a busy restaurant. To complete an order, **Character1** grabs an opaque **Object1** and fills it with **State1**. Then **Character2** grabs another opaque **Object2** and fills it with **State2**.” We then ask the LM to reason about one of the characters’ beliefs regarding the state of an object: “What does **Character1** believe **Object2** contains?” We analyze the LM’s ability to track characters’ beliefs in two distinct settings. (1) *No Visibility*, where both characters are unaware of each other’s actions, and (2) *Explicit Visibility* where explicit information about whether a character can/cannot observe the other’s actions is provided, e.g., “**Bob** can observe **Carla**’s actions. **Carla** cannot observe **Bob**’s actions.” We also provide general task instructions (e.g., answer unknown when a character is unaware); refer to Appendix A & C for the full prompt and additional dataset details. Our experiments analyze the Llama-3-70B-Instruct model in half-precision, using *NNsight* [Fiotto-Kaufman et al., 2025]. The model demonstrates a high behavioral performance on both the no-visibility and explicit-visibility settings, achieving accuracy of 95.7% and 99% respectively. For all subsequent experiments, we filter out samples that the model fails to answer correctly.

Causal Mediation Analysis

Our goal is to develop a mechanistic understanding of how Llama-3-70B-Instruct reasons about characters’ beliefs and answers related questions [Saphra and Wiegrefe, 2024]. A key method for conducting causal analysis is *interchange interventions* [Vig et al., 2020, Geiger et al., 2020, Finlayson et al., 2021], in which the LM is run on paired examples: an *original input* o and a *counterfactual input* c and certain internal activations in the LM run on the original are replaced with those computed from the counterfactual.

Drawing inspiration from existing literature [Vig et al., 2020, Meng et al., 2022, Wang et al., 2023], we begin our analysis by performing interchange interventions with counterfactuals that are identical to the original except for key input tokens. We trace the causal path from these key tokens to the final output. This is a type of *Causal Mediation Analysis* [Mueller et al., 2024]. Specifically, we construct a counterfactual dataset where o contains a question about the belief of a character not mentioned in the story, while c is identical except that the story includes the queried character. The expected outcome of this intervention is a change in the final output of o from *unknown* to a state token, such as **beer**. We conduct similar interchange interventions for object and state tokens (refer to Appendix D for more details).

Figure 3 presents the aggregated results of this experiment for the key input tokens **Character1**, **Object1**, and **State1**. The cells are color-coded to indicate the *interchange intervention accuracy*

[IIA; Geiger et al., 2022]. Even at this coarse level of analysis, 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 how information flows from an input token to the output but do not identify what that information is. A fact about transformers is that the input to the first layer contains input tokens and the output from the final layer contains the output token, but what is the information content of representations along the active causal path in between the input and output?

To answer this question, we turn to *Causal Abstraction* [Geiger et al., 2021, 2024]. We align the variables of a high-level causal model with the LM’s internal activations and verify the alignment by conducting targeted interchange interventions for each variable. Specifically, we perform aligned interchange interventions at both levels: interventions that target high-level causal variables and interventions that modify low-level features of the LM’s hidden activations. If the LM produces the same output as the high-level causal model under these aligned interventions, it provides evidence supporting the hypothesized causal model. The effect of these interventions is quantified using IIA, which measures the proportion of instances where intervened high-level causal model and low-level LM have the same output (refer to Appendix E for more details about the causal abstraction framework and Appendix F for the belief tracking causal model).

In addition to performing interchange interventions on entire residual stream vectors in LMs, we also intervene on specific subspaces to further localize causal variables. To identify the subspace encoding a particular variable, we employ the *Desiderata-based Component Masking* [De Cao et al., 2020, Davies et al., 2023, Prakash et al., 2024] technique, which learns a sparse binary mask over the internal activation space by maximizing the logit of the causal model output token. Specifically, we train a mask to select the singular vectors of the activation space that encode a high-level variable (see Appendix G for details).

4 Belief Tracking via Ordering IDs and Lookback Mechanisms

The LM solves the no visibility setting of the belief tracking task using three key mechanisms: *Ordering ID assignment*, *binding lookback*, and *answer lookback*. Figure 2a illustrates the hypothesized high-level causal model implemented by the LM, which we evaluate in the following subsections. The LM first assigns *ordering IDs* (OIs; , , ) to each **character**, **object**, and **state** in the story that encode their order of appearance (e.g., the second character **Carla** is assigned ). These OIs are used in two lookback mechanisms. (i) **Binding lookback**: *Address* copies of each character OI () and object OI () are placed alongside their corresponding state OI *payload* () in the residual stream of each state token, binding together each character-object-state triple. When the model is asked about the belief of a specific character about a specific object, it moves *pointer* copies of the corresponding OIs (, ) to the final token’s residual stream. These pointers are dereferenced, bringing the correct state OI into the final token residual stream. (ii) **Answer lookback**: An *address* copy of the state OI () is alongside the state token *payload* () in the residual stream of the correct state token, while a *pointer* copy () is moved to the final token residual stream via the binding lookback. The pointer is dereferenced, bringing the answer state token payload into the final token residual stream, which is predicted as the final output. Refer to Appendix F for pseudocode defining the causal model for the belief tracking task. In Appendices M and L, we show parts of our analysis generalizing to the Llama-3.1-405B-Instruct model and the BigToM dataset [Gandhi et al., 2024].

4.1 Ordering ID Assignment

The LM assigns an Ordering ID (OI; Dai et al. 2024) to the character, object, and state tokens. These OIs, encoded in a low-rank subspace of the internal activation, serve as a reference that indicates whether an entity is the first or second of its type independent of its token value. For example, in Fig. 2a, **Bob** is assigned the first character OI, while **Carla** receives the second. In the subsequent subsections and Appendices H & I, we validate the presence of OIs through multiple experiments, where intervening on tokens with identical token values but different OIs alters the model’s internal

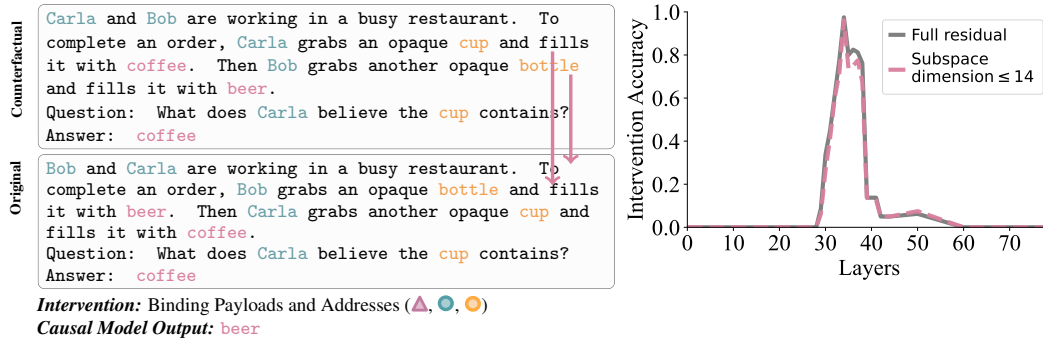


Figure 4: **Binding lookback payload and address:** We intervened on both the high-level causal model and the LM running on the original story, modifying their variables and internal activations respectively, to match those from a counterfactual scenario. In the causal model intervention, we update the addresses (character and object OIs; \triangle and \circ) and the payloads (state OIs; \triangle). This causes the binding lookback mechanism to attend to and retrieve the state OI corresponding to the alternate state token, which is then dereferenced by answer lookback to yield the alternate state token (e.g., *beer* instead of *coffee*). In the LM interchange intervention, modifying the residual stream at the state token results in identical outputs between layers 33 and 38. This confirms our hypothesis that both the address and payload information are represented in the residual stream of state tokens.

computation, leading to systematic changes in the final output predicted by our high-level causal model. The LM then uses these OIs as building blocks, feeding them into lookback mechanisms to track and retrieve beliefs.

4.2 Uncovering the Binding Lookback Mechanism

The *Binding Lookback* is the first operation applied to these OIs. The character and object OIs, serving as the source information, are each copied twice. One copy, referred to as the address, is placed in the residual stream of the state token (recalled token), alongside the state OI as the payload to transfer. The other copy, referred to as the pointer, is moved in the residual stream of the final token (lookback token). These pointer and address copies are then used to form the QK-circuit at the lookback token, which dereferences the state OI payload, transferring it from the state token to the final token. See Fig.2a (i) for a schematic of this lookback and see Fig.1 for the general mechanism.

The Hypothesized Address and Payload. In our first experiment, we localize the address copies of the character and object OIs and the state OI payload to the residual stream of the state token (recalled token, Fig. 2a). We sampled a counterfactual dataset where each example consists of an original input \mathbf{o} with an answer that isn't *unknown* and a counterfactual input \mathbf{c} where the character, object, and state tokens are identical, except the ordering of the two sentences is swapped while the question remains unchanged, as illustrated in Fig. 4. The expected outcome predicted by our high-level causal model under intervention is the other state token from the original example, e.g., *beer*, because reversing the address and payload values without changing the pointer flips the output.

Testing Address and Payload Hypothesis. We perform an interchange intervention experiment layer-by-layer, where we replace the residual stream vectors at 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. 4, the strongest alignment occurs between layers 33 and 38, supporting our hypothesis that the state token's residual stream contains both the address information (character and object OIs) and the payload information (state OI). These components are subsequently used to form a QK-circuit between the pointer at the lookback token and the address at the other state token and OV-circuit that retrieves its state OI as the payload.

Localizing the Source Information As shown in Fig. 2a, the source information is copied as both the address and the pointer at different token positions. To localize the source information, we conduct intervention experiments 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} .

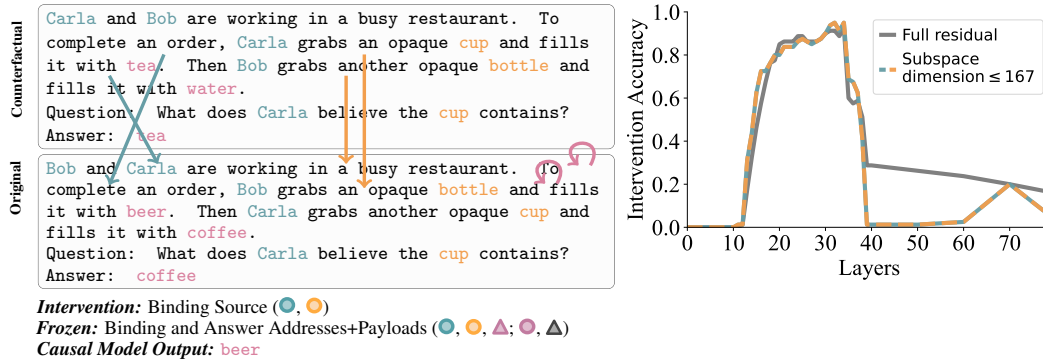


Figure 5: **Source Information of Binding lookback:** We run the causal model and the LM on an original story, then update their variables and activations, respectively, to the values they would take on for a counterfactual story with swapped characters and objects and new states. The interchange intervention on the high-level causal model swaps the sources of the binding lookback (character and object OIs; ●, ●) while freezing the addresses and payloads of the binding lookback (character, object, and state OIs; ●, ●, ▲) and the answer lookback (state OI and token; ▲, ▲). By altering the sources, but freezing the addresses and payloads, only the pointer is changed so the binding lookback retrieves the other state OI which is dereferenced by the answer lookback to the other state token (e.g., beer instead of coffee). We perform the same interchange intervention on the LM and measure the agreement with the intervened causal model. Our results localize the source to the character and object token residual streams between layers 20 and 34.

218 With this dataset, an interchange intervention on the high-level causal model that targets the source
 219 information will have downstream effects on both the address and the pointer, so no change in output
 220 occurs. However, if we additionally freeze the payloads and addresses, the causal model outputs the
 221 other state token, e.g., beer in Fig. 5, due to the mismatch between address and pointer.

222 In the LM, we interchange the residual streams of the character and object tokens while keeping the
 223 residual stream of the state token fixed. When the output of the intervened LM aligns with that of the
 224 intervened causal model, it indicates that the QK-circuit at the final token is attending to the alternate
 225 state token. As shown in Fig. 5, the second experiment reveals alignment between layers 20 and 34.
 226 This suggests that source information—specifically, the character and object OIs—is represented in
 227 their respective token residual streams within this layer range.

228 We provide more experimental results in Appendix H where we show in Fig. 12 that freezing the
 229 residual stream of the state token is necessary. In sum, these results not only provide evidence for the
 230 presence of source information but also establish its transfer to the recalled and lookback tokens as
 231 addresses and pointers, respectively.

232 **Localizing the Pointer Information** The pointer copies of the character and object OI are first
 233 formed at the character and object tokens in the question before being moved again to the final token
 234 for dereferencing (see Appendix I for experiments and more details).

235 4.3 Uncovering the Answer Lookback Mechanism

236 The LM answers the question using the *Answer Lookback*. The state OI of the correct answer serves
 237 as the source information, which is copied into two instances. One instance, the address copy of
 238 the state OI, is in the residual stream of the state token (recalled token) with the state token itself
 239 as the payload. The other instance, the pointer copy of the state OI, is transferred to the residual
 240 stream of the final token (lookback token) as the payload of the binding lookback. This pointer is
 241 then dereferenced, bringing the state token as the payload into the residual stream of the final token,
 242 which is predicted as the final output.

243 **Localizing the Pointer Information** We first localize the pointer of the answer lookback, which is
 244 the payload of the binding lookback. To do this, we conduct an interchange intervention experiment
 245 where the residual vectors at the final token position in the original run are replaced with those from
 246 the counterfactual run, one layer at a time. The counterfactual inputs have swapped objects and
 247 characters and randomly sampled states. If the answer pointer is targeted for intervention in the
 248 high-level causal model, the output is the other state in the original input, e.g., beer. As shown

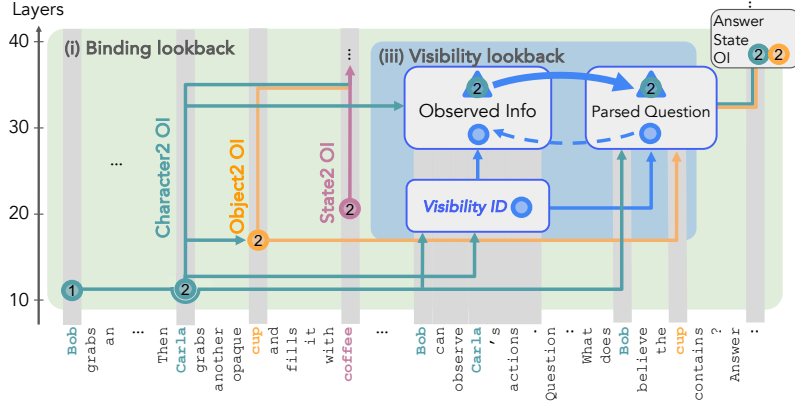


Figure 6: **Visibility Lookback** When one character (the observing character) can see another (the observed character), the LM assigns a visibility ID (●) to the visibility sentence (where this relation is defined). An address copy of this visibility ID remains in the visibility sentence’s residual stream. A pointer copy of the visibility ID is transferred to the subsequent tokens’ residual stream (lookback tokens). During processing, the model dereferences this pointer through a QK-circuit, bringing forward the payload (▲). Based on initial evidence, this payload contains the observed character’s OI(●). Refer to Appendix J for more 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.

in Fig. 2b, alignment begins at layer 34, indicating that this layer contains pointer information, in low-rank subspace, which remains causally relevant until layer 52.

Localizing the Payload To determine where the model uses the state OI pointer to retrieve the state token, we use the same interchange intervention experiment. However, if the answer payload is targeted for intervention in the high-level causal model, the output is the correct state token from the counterfactual example, e.g., *tea*, rather than the state token from the original example, as illustrated in Fig. 2b. The alignment occurs after layer 56, indicating that the model retrieves the correct state token (payload) into the final token’s residual stream by 56, which is used to generate the final output.

5 Impact of Visibility Conditions on Belief Tracking Mechanism

In the previous section, we 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 provided with additional information that one of the characters (*observing*) can observe the actions of others (*observed*). We hypothesize that the LM employs another lookback mechanism, which we refer to as the *Visibility Lookback*, to incorporate information about the observed character.

As illustrated in Fig. 6, we hypothesize that the LM first generates a *Visibility ID* at the residual stream of the visibility sentence, serving as the source 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 streams of the subsequent tokens, which are the lookback tokens. Then LM forms a QK-circuit at the lookback tokens and dereferences the visibility ID pointer to bring the payload.

Although we were unable to determine the exact semantics of the payload in this lookback, we speculate that it encodes the observed character’s OI. We propose the existence of another lookback, where the story sentence associated with the observed character serves as the source, and its payload encodes information about the observed character. That payload information, encoding observed character’s OI, is then retrieved by the lookback tokens of the Visibility lookback, which contributes to the queried character’s enhanced awareness (see Appendix J for more details).

5.1 Uncovering the Visibility Lookback Mechanism

Localizing the Source Information To localize the source information, we conduct an interchange intervention experiment where the counterfactual is a different story with altered visibility information.

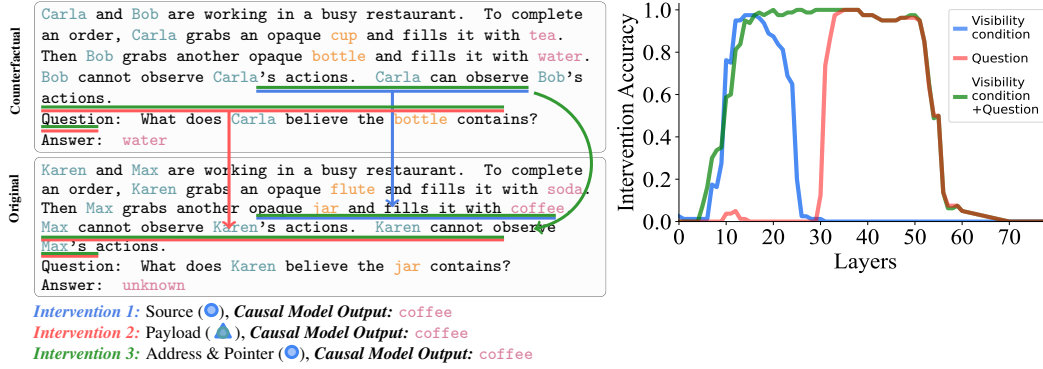


Figure 7: **Visibility Lookback:** We conduct three interchange intervention experiments to support the Visibility Lookback hypothesis: (1) *Source Alignment*: We align the source information (●) by intervening on the visibility sentence-replacing it with its representation from a counterfactual run where the visibility sentence causes the queried character to become aware of the queried object’s contents. We observe that source information aligns between layers 10 and 23. (2) *Payload Alignment* (▲): 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.

278 In the original example, the first character cannot observe the second character’s actions, whereas in
 279 the counterfactual example, the first character can observe them (Fig. 7). The causal model outcome
 280 of this intervention is a change in the final output of the original run from “unknown” to the state token
 281 associated with the queried object. The interchange intervention is executed on visibility sentence
 282 tokens. As shown in Fig. 7 (— line), alignment occurs between layers 10 and 23, indicating that the
 283 visibility ID remains encoded in the visibility sentence until layer 23, after which it is duplicated into
 284 address and pointer copies on visibility sentence and subsequent tokens respectively.

285 **Localizing the Payload** To localize the payload information, we use the same counterfactual
 286 dataset. However, instead of intervening on the source or recalled tokens, we intervene on the
 287 lookback tokens, specifically the question and answer tokens. As in the previous experiment, we
 288 replace the residual vectors of these tokens in the original run with those from the counterfactual run.
 289 As shown in Fig. 7 (— line), alignment occurs only after layer 31, indicating that the information
 290 enhancing the queried character’s awareness is present in the lookback tokens only after this layer.

291 **Localizing the Address and Pointer** The previous two experiments suggest the presence of a
 292 lookback mechanism, as there is no signal indicating that the source or payload has been formed
 293 between layers 24 and 31. We hypothesize that this lack of signal is due to a mismatch between the
 294 address and pointer information. Specifically, when intervening only on the recalled token after layer
 295 25, the pointer is not updated, whereas intervening only on the lookback tokens leaves the address
 296 unaltered, leading to the mismatch. To test this hypothesis, we conduct another intervention using
 297 the same counterfactual dataset, where we intervene on the residual vectors of both the recalled and
 298 lookback tokens. As shown in Fig. 7 (— line), alignment occurs after layer 10 and remains stable,
 299 supporting our hypothesis. This intervention replaces both the address and pointer copies of the
 300 visibility IDs, enabling the LM to form a QK-circuit and retrieve the payload.

301 6 Conclusion

302 Through a series of desiderata-based patching experiments, we have mapped the mechanisms un-
 303 derlying the processing of partial knowledge and false beliefs in a set of simple stories. We are
 304 surprised by the pervasive appearance of a single recurring computational pattern: the lookback,
 305 which resembles a pointer dereference inside a transformer. The LMs use a combination of several
 306 lookbacks to reason about nontrivial visibility and belief states. Our improved understanding of these
 307 fundamental computations gives us optimism that it may be possible to fully reveal the algorithms
 308 underlying not only Theory of Mind, but also other forms of reasoning in LMs.

References

- G. Alain. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.
- S. Baron-Cohen, A. M. Leslie, and U. Frith. Does the autistic child have a “theory of mind”? *Cognition*, 21(1):37–46, 1985.
- Y. Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.
- M. Bortoletto, C. Ruhdorfer, L. Shi, and A. Bulling. Benchmarking mental state representations in language models. *arXiv preprint arXiv:2406.17513*, 2024.
- C. Chan, C. Jiayang, Y. Yim, Z. Deng, W. Fan, H. Li, X. Liu, H. Zhang, W. Wang, and Y. Song. Negotiationtom: A benchmark for stress-testing machine theory of mind on negotiation surrounding. *arXiv preprint arXiv:2404.13627*, 2024.
- Q. Dai, B. Heinzerling, and K. Inui. Representational analysis of binding in language models. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17468–17493, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.967. URL <https://aclanthology.org/2024.emnlp-main.967/>.
- X. Davies, M. Nadeau, N. Prakash, T. R. Shaham, and D. Bau. Discovering variable binding circuitry with desiderata, 2023. URL <https://arxiv.org/abs/2307.03637>.
- N. De Cao, M. S. Schlichtkrull, W. Aziz, and I. Titov. How do decisions emerge across layers in neural models? interpretation with differentiable masking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3243–3255, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.262. URL <https://aclanthology.org/2020.emnlp-main.262>.
- D. C. Dennett. *The Intentional Stance*. MIT Press, 1981.
- N. Elhage, N. Nanda, C. Olsson, T. Henighan, N. Joseph, B. Mann, A. Askell, Y. Bai, A. Chen, T. Conerly, N. DasSarma, D. Drain, D. Ganguli, Z. Hatfield-Dodds, D. Hernandez, A. Jones, J. Kernion, L. Lovitt, K. Ndousse, D. Amodei, T. Brown, J. Clark, J. Kaplan, S. McCandlish, and C. Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. <https://transformer-circuits.pub/2021/framework/index.html>.
- J. Feng and J. Steinhardt. How do language models bind entities in context? *arXiv preprint arXiv:2310.17191*, 2023.
- J. Feng, S. Russell, and J. Steinhardt. Monitoring latent world states in language models with propositional probes. *CoRR*, abs/2406.19501, 2024. doi: 10.48550/ARXIV.2406.19501. URL <https://doi.org/10.48550/arXiv.2406.19501>.
- M. Finlayson, A. Mueller, S. Gehrmann, S. M. Shieber, T. Linzen, and Y. Belinkov. Causal analysis of syntactic agreement mechanisms in neural language models. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 1828–1843. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.ACL-LONG.144. URL <https://doi.org/10.18653/v1/2021.acl-long.144>.
- J. F. Fiotto-Kaufman, A. R. Loftus, E. Todd, J. Brinkmann, K. Pal, D. Troitskii, M. Ripa, A. Belfki, C. Rager, C. Juang, A. Mueller, S. Marks, A. S. Sharma, F. Lucchetti, N. Prakash, C. E. Brodley, A. Guha, J. Bell, B. C. Wallace, and D. Bau. NNsight and NDIF: Democratizing access to open-weight foundation model internals. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=MxbEiFRf39>.

- 356 K. Gandhi, J.-P. Fränken, T. Gerstenberg, and N. Goodman. Understanding social reasoning in
357 language models with language models. *Advances in Neural Information Processing Systems*, 36,
358 2024.
- 359 A. Geiger, K. Richardson, and C. Potts. Neural natural language inference models partially em-
360 bed theories of lexical entailment and negation. In A. Alishahi, Y. Belinkov, G. Chrupała,
361 D. Hupkes, Y. Pinter, and H. Sajjad, editors, *Proceedings of the Third BlackboxNLP Workshop*
362 *on Analyzing and Interpreting Neural Networks for NLP*, pages 163–173, Online, Nov. 2020.
363 Association for Computational Linguistics. doi: 10.18653/v1/2020.blackboxnlp-1.16. URL
364 <https://aclanthology.org/2020.blackboxnlp-1.16>.
- 365 A. Geiger, H. Lu, T. Icard, and C. Potts. Causal abstractions of neural networks. In
366 M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, editors,
367 *Advances in Neural Information Processing Systems 34: Annual Conference on Neu-*
368 *ral Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*,
369 pages 9574–9586, 2021. URL [https://proceedings.neurips.cc/paper/2021/hash/](https://proceedings.neurips.cc/paper/2021/hash/4f5c422f4d49a5a807eda27434231040-Abstract.html)
370 [4f5c422f4d49a5a807eda27434231040-Abstract.html](https://proceedings.neurips.cc/paper/2021/hash/4f5c422f4d49a5a807eda27434231040-Abstract.html).
- 371 A. Geiger, Z. Wu, H. Lu, J. Rozner, E. Kreiss, T. Icard, N. Goodman, and C. Potts. Inducing causal
372 structure for interpretable neural networks. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari,
373 G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine*
374 *Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 7324–7338. PMLR,
375 17–23 Jul 2022. URL <https://proceedings.mlr.press/v162/geiger22a.html>.
- 376 A. Geiger, D. Ibeling, A. Zur, M. Chaudhary, S. Chauhan, J. Huang, A. Arora, Z. Wu, N. Goodman,
377 C. Potts, and T. Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability,
378 2024. URL <https://arxiv.org/abs/2301.04709>.
- 379 M. Geva, J. Bastings, K. Filippova, and A. Globerson. Dissecting recall of factual associations in
380 auto-regressive language models, 2023. URL <https://arxiv.org/abs/2304.14767>.
- 381 A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur,
382 A. Schelten, A. Vaughan, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Sra-
383 vankumar, A. Korenev, A. Hinsvark, A. Rao, A. Zhang, A. Rodriguez, A. Gregerson, A. Spataru,
384 B. Roziere, B. Biron, B. Tang, B. Chern, C. Caucheteux, C. Nayak, C. Bi, C. Marra, C. McConnell,
385 C. Keller, C. Touret, C. Wu, C. Wong, C. C. Ferrer, C. Nikolaidis, D. Allonsius, D. Song, D. Pintz,
386 D. Livshits, D. Wyatt, D. Esiobu, D. Choudhary, D. Mahajan, D. Garcia-Olano, D. Perino, D. Hup-
387 kes, E. Lakomkin, E. AlBadawy, E. Lobanova, E. Dinan, E. M. Smith, F. Radenovic, F. Guzmán,
388 F. Zhang, G. Synnaeve, G. Lee, G. L. Anderson, G. Thattai, G. Nail, G. Mialon, G. Pang, G. Cu-
389 curell, H. Nguyen, H. Korevaar, H. Xu, H. Touvron, I. Zarov, I. A. Ibarra, I. Kloumann, I. Misra,
390 I. Evtimov, J. Zhang, J. Copet, J. Lee, J. Geffert, J. Vranes, J. Park, J. Mahadeokar, J. Shah,
391 J. van der Linde, J. Billock, J. Hong, J. Lee, J. Fu, J. Chi, J. Huang, J. Liu, J. Wang, J. Yu, J. Bitton,
392 J. Spisak, J. Park, J. Rocca, J. Johnston, J. Saxe, J. Jia, K. V. Alwala, K. Prasad, K. Upasani, K. Plaw-
393 iak, K. Li, K. Heafield, K. Stone, K. El-Arini, K. Iyer, K. Malik, K. Chiu, K. Bhalla, K. Lakhota,
394 L. Rantala-Yeary, L. van der Maaten, L. Chen, L. Tan, L. Jenkins, L. Martin, L. Madaan, L. Malo,
395 L. Blecher, L. Landzaat, L. de Oliveira, M. Muzzi, M. Pasupuleti, M. Singh, M. Paluri, M. Kar-
396 das, M. Tsimpoukelli, M. Oldham, M. Rita, M. Pavlova, M. Kambadur, M. Lewis, M. Si, M. K.
397 Singh, M. Hassan, N. Goyal, N. Torabi, N. Bashlykov, N. Bogoychev, N. Chatterji, N. Zhang,
398 O. Duchenne, O. Çelebi, P. Alrassy, P. Zhang, P. Li, P. Vasic, P. Weng, P. Bhargava, P. Dubal, P. Kr-
399 ishnan, P. S. Koura, P. Xu, Q. He, Q. Dong, R. Srinivasan, R. Ganapathy, R. Calderer, R. S. Cabral,
400 R. Stojnic, R. Raileanu, R. Maheswari, R. Girdhar, R. Patel, R. Sauvestre, R. Polidoro, R. Sumbaly,
401 R. Taylor, R. Silva, R. Hou, R. Wang, S. Hosseini, S. Chennabasappa, S. Singh, S. Bell, S. S. Kim,
402 S. Edunov, S. Nie, S. Narang, S. Raparthy, S. Shen, S. Wan, S. Bhosale, S. Zhang, S. Vandenhende,
403 S. Batra, S. Whitman, S. Sootla, S. Collot, S. Gururangan, S. Borodinsky, T. Herman, T. Fowler,
404 T. Sheasha, T. Georgiou, T. Scialom, T. Speckbacher, T. Mihaylov, T. Xiao, U. Karn, V. Goswami,
405 V. Gupta, V. Ramanathan, V. Kerkez, V. Gonguet, V. Do, V. Vogeti, V. Albiero, V. Petrovic, W. Chu,
406 W. Xiong, W. Fu, W. Meers, X. Martinet, X. Wang, X. Wang, X. E. Tan, X. Xia, X. Xie, X. Jia,
407 X. Wang, Y. Goldschlag, Y. Gaur, Y. Babaei, Y. Wen, Y. Song, Y. Zhang, Y. Li, Y. Mao, Z. D.
408 Coudert, Z. Yan, Z. Chen, Z. Papakipos, A. Singh, A. Srivastava, A. Jain, A. Kelsey, A. Shajnfeld,
409 A. Gangidi, A. Victoria, A. Goldstand, A. Menon, A. Sharma, A. Boesenberg, A. Baevski, A. Fein-
410 stein, A. Kallet, A. Sangani, A. Teo, A. Yunus, A. Lupu, A. Alvarado, A. Caples, A. Gu, A. Ho,

- 411 A. Poulton, A. Ryan, A. Ramchandani, A. Dong, A. Franco, A. Goyal, A. Saraf, A. Chowdhury,
412 A. Gabriel, A. Bharambe, A. Eisenman, A. Yazdan, B. James, B. Maurer, B. Leonhardi, B. Huang,
413 B. Loyd, B. D. Paola, B. Paranjape, B. Liu, B. Wu, B. Ni, B. Hancock, B. Wasti, B. Spence,
414 B. Stojkovic, B. Gamido, B. Montalvo, C. Parker, C. Burton, C. Mejia, C. Liu, C. Wang, C. Kim,
415 C. Zhou, C. Hu, C.-H. Chu, C. Cai, C. Tindal, C. Feichtenhofer, C. Gao, D. Civin, D. Beaty,
416 D. Kreymer, D. Li, D. Adkins, D. Xu, D. Testuggine, D. David, D. Parikh, D. Liskovich, D. Foss,
417 D. Wang, D. Le, D. Holland, E. Dowling, E. Jamil, E. Montgomery, E. Presani, E. Hahn, E. Wood,
418 E.-T. Le, E. Brinkman, E. Arcaute, E. Dunbar, E. Smothers, F. Sun, F. Kreuk, F. Tian, F. Kokkinos,
419 F. Ozgenel, F. Caggioni, F. Kanayet, F. Seide, G. M. Florez, G. Schwarz, G. Badeer, G. Sweet,
420 G. Halpern, G. Herman, G. Sizov, Guangyi, Zhang, G. Lakshminarayanan, H. Inan, H. Shojanazeri,
421 H. Zou, H. Wang, H. Zha, H. Habeeb, H. Rudolph, H. Suk, H. Aspegren, H. Goldman, H. Zhan,
422 I. Damla, I. Molybog, I. Tufanov, I. Leontiadis, I.-E. Veliche, I. Gat, J. Weissman, J. Geboski,
423 J. Kohli, J. Lam, J. Asher, J.-B. Gaya, J. Marcus, J. Tang, J. Chan, J. Zhen, J. Reizenstein, J. Teboul,
424 J. Zhong, J. Jin, J. Yang, J. Cummings, J. Carvill, J. Shepard, J. McPhie, J. Torres, J. Ginsburg,
425 J. Wang, K. Wu, K. H. U, K. Saxena, K. Khandelwal, K. Zand, K. Matosich, K. Veeraraghavan,
426 K. Michelena, K. Li, K. Jagadeesh, K. Huang, K. Chawla, K. Huang, L. Chen, L. Garg, L. A.
427 L. Silva, L. Bell, L. Zhang, L. Guo, L. Yu, L. Moshkovich, L. Wehrstedt, M. Khabsa, M. Avalani,
428 M. Bhatt, M. Mankus, M. Hasson, M. Lennie, M. Reso, M. Groshev, M. Naumov, M. Lathi,
429 M. Keneally, M. Liu, M. L. Seltzer, M. Valko, M. Restrepo, M. Patel, M. Vyatskov, M. Samvelyan,
430 M. Clark, M. Macey, M. Wang, M. J. Hermoso, M. Metanat, M. Rastegari, M. Bansal, N. San-
431 thanam, N. Parks, N. White, N. Bawa, N. Singhal, N. Egebo, N. Usunier, N. Mehta, N. P. Laptev,
432 N. Dong, N. Cheng, O. Chernoguz, O. Hart, O. Salpekar, O. Kalinli, P. Kent, P. Parekh, P. Saab,
433 P. Balaji, P. Rittner, P. Bontrager, P. Roux, P. Dollar, P. Zvyagina, P. Ratanchandani, P. Yuvraj,
434 Q. Liang, R. Alao, R. Rodriguez, R. Ayub, R. Murthy, R. Nayani, R. Mitra, R. Parthasarathy,
435 R. Li, R. Hogan, R. Battey, R. Wang, R. Howes, R. Rinott, S. Mehta, S. Siby, S. J. Bondu,
436 S. Datta, S. Chugh, S. Hunt, S. Dhillon, S. Sidorov, S. Pan, S. Mahajan, S. Verma, S. Yamamoto,
437 S. Ramaswamy, S. Lindsay, S. Lindsay, S. Feng, S. Lin, S. C. Zha, S. Patil, S. Shankar, S. Zhang,
438 S. Zhang, S. Wang, S. Agarwal, S. Sajuyigbe, S. Chintala, S. Max, S. Chen, S. Kehoe, S. Satterfield,
439 S. Govindaprasad, S. Gupta, S. Deng, S. Cho, S. Virk, S. Subramanian, S. Choudhury, S. Goldman,
440 T. Remez, T. Glaser, T. Best, T. Koehler, T. Robinson, T. Li, T. Zhang, T. Matthews, T. Chou,
441 T. Shaked, V. Vontimitta, V. Ajayi, V. Montanez, V. Mohan, V. S. Kumar, V. Mangla, V. Ionescu,
442 V. Poenaru, V. T. Mihailescu, V. Ivanov, W. Li, W. Wang, W. Jiang, W. Bouaziz, W. Constable,
443 X. Tang, X. Wu, X. Wang, X. Wu, X. Gao, Y. Kleinman, Y. Chen, Y. Hu, Y. Jia, Y. Qi, Y. Li,
444 Y. Zhang, Y. Zhang, Y. Adi, Y. Nam, Yu, Wang, Y. Zhao, Y. Hao, Y. Qian, Y. Li, Y. He, Z. Rait,
445 Z. DeVito, Z. Rosnbrick, Z. Wen, Z. Yang, Z. Zhao, and Z. Ma. The llama 3 herd of models, 2024.
446 URL <https://arxiv.org/abs/2407.21783>.
- 447 H. Gweon, J. Fan, and B. Kim. Socially intelligent machines that learn from humans and help humans
448 learn. *Philosophical Transactions of the Royal Society A*, 381(2251):20220048, 2023.
- 449 D. A. Herrmann and B. A. Levinstein. Standards for belief representations in llms. *arXiv preprint*
450 *arXiv:2405.21030*, 2024.
- 451 G. Hou, W. Zhang, Y. Shen, L. Wu, and W. Lu. TimeToM: Temporal space is the key to unlocking
452 the door of large language models’ theory-of-mind. In L.-W. Ku, A. Martins, and V. Srikumar,
453 editors, *Findings of the Association for Computational Linguistics ACL 2024*, pages 11532–11547,
454 Bangkok, Thailand and virtual meeting, Aug. 2024. Association for Computational Linguistics.
455 URL <https://aclanthology.org/2024.findings-acl.685>.
- 456 J. Hu, F. Sosa, and T. Ullman. Re-evaluating theory of mind evaluation in large language models.
457 *arXiv preprint arXiv:2502.21098*, 2025.
- 458 C. Jin, Y. Wu, J. Cao, J. Xiang, Y.-L. Kuo, Z. Hu, T. Ullman, A. Torralba, J. Tenenbaum, and T. Shu.
459 MMTOM-QA: Multimodal theory of mind question answering. In L.-W. Ku, A. Martins, and
460 V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational*
461 *Linguistics (Volume 1: Long Papers)*, pages 16077–16102, Bangkok, Thailand, Aug. 2024. Associa-
462 tion for Computational Linguistics. URL <https://aclanthology.org/2024.acl-long.851>.
- 463 H. Kim, M. Sclar, X. Zhou, R. Bras, G. Kim, Y. Choi, and M. Sap. FANTOM: A benchmark for
464 stress-testing machine theory of mind in interactions. In H. Bouamor, J. Pino, and K. Bali, editors,
465 *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages

14397–14413, Singapore, Dec. 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.890. URL <https://aclanthology.org/2023.emnlp-main.890/>.

H. Kim, M. Sclar, X. Zhou, R. L. Bras, G. Kim, Y. Choi, and M. Sap. Fantom: A benchmark for stress-testing machine theory of mind in interactions. *arXiv preprint arXiv:2310.15421*, 2023b.

N. Kim and S. Schuster. Entity tracking in language models. *arXiv preprint arXiv:2305.02363*, 2023.

Y. Kim and S. S. Sundar. Anthropomorphism of computers: Is it mindful or mindless? *Computers in Human Behavior*, 28(1):241–250, 2012.

M. Kosinski. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45), Oct. 2024. ISSN 1091-6490. doi: 10.1073/pnas.2405460121. URL <http://dx.doi.org/10.1073/pnas.2405460121>.

M. Le, Y.-L. Boureau, and M. Nickel. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5872–5877, 2019.

B. Z. Li, M. Nye, and J. Andreas. Implicit representations of meaning in neural language models. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1813–1827, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.143. URL <https://aclanthology.org/2021.acl-long.143>.

K. Li, O. Patel, F. Viégas, H. Pfister, and M. Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36, 2024.

T. Lieberum, M. Rahtz, J. Kramár, N. Nanda, G. Irving, R. Shah, and V. Mikulik. Does circuit analysis interpretability scale? evidence from multiple choice capabilities in chinchilla, 2023. URL <https://arxiv.org/abs/2307.09458>.

K. Meng, D. Bau, A. Andonian, and Y. Belinkov. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 36, 2022.

S. R. Moghaddam and C. J. Honey. Boosting theory-of-mind performance in large language models via prompting. *arXiv preprint arXiv:2304.11490*, 2023.

A. Mueller, J. Brinkmann, M. Li, S. Marks, K. Pal, N. Prakash, C. Rager, A. Sankaranarayanan, A. S. Sharma, J. Sun, E. Todd, D. Bau, and Y. Belinkov. The quest for the right mediator: A history, survey, and theoretical grounding of causal interpretability, 2024. URL <https://arxiv.org/abs/2408.01416>.

C. Olsson, N. Elhage, N. Nanda, N. Joseph, N. DasSarma, T. Henighan, B. Mann, A. Askell, Y. Bai, A. Chen, T. Conerly, D. Drain, D. Ganguli, Z. Hatfield-Dodds, D. Hernandez, S. Johnston, A. Jones, J. Kernion, L. Lovitt, K. Ndousse, D. Amodei, T. Brown, J. Clark, J. Kaplan, S. McCandlish, and C. Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. <https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html>.

N. Prakash, T. R. Shaham, T. Haklay, Y. Belinkov, and D. Bau. Fine-tuning enhances existing mechanisms: A case study on entity tracking. In *Proceedings of the 2024 International Conference on Learning Representations*, 2024. arXiv:2402.14811.

D. Premack and G. Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526, 1978. doi: 10.1017/S0140525X00076512.

C. Riedl, Y. J. Kim, P. Gupta, T. W. Malone, and A. W. Woolley. Quantifying collective intelligence in human groups. *Proceedings of the National Academy of Sciences*, 118(21):e2005737118, 2021.

N. Rimskey, N. Gabrieli, J. Schulz, M. Tong, E. Hubinger, and A. M. Turner. Steering llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2023.

- 514 N. Saphra and S. Wiegreffe. Mechanistic? In Y. Belinkov, N. Kim, J. Jumelet, H. Mohebbi,
515 A. Mueller, and H. Chen, editors, *Proceedings of the 7th BlackboxNLP Workshop: Analyzing
516 and Interpreting Neural Networks for NLP*, pages 480–498, Miami, Florida, US, Nov. 2024.
517 Association for Computational Linguistics. doi: 10.18653/v1/2024.blackboxnlp-1.30. URL
518 <https://aclanthology.org/2024.blackboxnlp-1.30/>.
- 519 M. Sclar, S. Kumar, P. West, A. Suhr, Y. Choi, and Y. Tsvetkov. Minding language models’(lack of)
520 theory of mind: A plug-and-play multi-character belief tracker. *arXiv preprint arXiv:2306.00924*,
521 2023.
- 522 M. Sclar, J. Yu, M. Fazel-Zarandi, Y. Tsvetkov, Y. Bisk, Y. Choi, and A. Celikyilmaz. Explore theory
523 of mind: Program-guided adversarial data generation for theory of mind reasoning. *ICLR*, 2025.
- 524 N. Shapira, G. Zwirn, and Y. Goldberg. How well do large language models perform on faux
525 pas tests? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages
526 10438–10451, Toronto, Canada, July 2023. Association for Computational Linguistics. URL
527 <https://aclanthology.org/2023.findings-acl.663>.
- 528 N. Shapira, M. Levy, S. H. Alavi, X. Zhou, Y. Choi, Y. Goldberg, M. Sap, and V. Shwartz. Clever hans
529 or neural theory of mind? stress testing social reasoning in large language models. In Y. Graham and
530 M. Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association
531 for Computational Linguistics (Volume 1: Long Papers)*, pages 2257–2273, St. Julian’s, Malta,
532 Mar. 2024. Association for Computational Linguistics. URL [https://aclanthology.org/](https://aclanthology.org/2024.eacl-long.138)
533 [2024.eacl-long.138](https://aclanthology.org/2024.eacl-long.138).
- 534 P. Smolensky. Neural and conceptual interpretation of PDP models. In J. L. McClelland, D. E.
535 Rumelhart, and the PDP Research Group, editors, *Parallel Distributed Processing: Explorations in
536 the Microstructure of Cognition: Psychological and Biological Models*, volume 2, pages 390–431.
537 MIT Press, 1986.
- 538 J. W. Strachan, D. Albergo, G. Borghini, O. Pansardi, E. Scaliti, S. Gupta, K. Saxena, A. Rufo,
539 S. Panzeri, G. Manzi, et al. Testing theory of mind in large language models and humans. *Nature
540 Human Behaviour*, 8(7):1285–1295, 2024a.
- 541 J. W. A. Strachan, D. Albergo, G. Borghini, O. Pansardi, E. Scaliti, S. Gupta, K. Saxena, A. Rufo,
542 S. Panzeri, G. Manzi, M. S. A. Graziano, and C. Becchio. Testing theory of mind in large language
543 models and humans. *Nature Human Behaviour*, 8(7):1285–1295, Jul 2024b. ISSN 2397-3374. doi:
544 [10.1038/s41562-024-01882-z](https://doi.org/10.1038/s41562-024-01882-z). URL <https://doi.org/10.1038/s41562-024-01882-z>.
- 545 W. Street, J. O. Siy, G. Keeling, A. Baranes, B. Barnett, M. McKibben, T. Kanyere, A. Lentz, B. A.
546 y Arcas, and R. I. M. Dunbar. Llms achieve adult human performance on higher-order theory of
547 mind tasks, 2024. URL <https://arxiv.org/abs/2405.18870>.
- 548 J. Vig, S. Gehrmann, Y. Belinkov, S. Qian, D. Nevo, Y. Singer, and S. Shieber. Investigating gender
549 bias in language models using causal mediation analysis. In H. Larochelle, M. Ranzato, R. Hadsell,
550 M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33,
551 pages 12388–12401. Curran Associates, Inc., 2020. URL [https://proceedings.neurips.cc/](https://proceedings.neurips.cc/paper_files/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf)
552 [paper_files/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf).
- 553 K. R. Wang, A. Variengien, A. Conmy, B. Shlegeris, and J. Steinhardt. Interpretability in the wild: a
554 circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference
555 on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.
556 URL <https://openreview.net/forum?id=NpsVSN6o4ul>.
- 557 A. Wilf, S. Lee, P. P. Liang, and L.-P. Morency. Think twice: Perspective-taking improves large
558 language models’ theory-of-mind capabilities. In L.-W. Ku, A. Martins, and V. Srikumar, editors,
559 *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume
560 1: Long Papers)*, pages 8292–8308, Bangkok, Thailand, Aug. 2024. Association for Computational
561 Linguistics. URL <https://aclanthology.org/2024.acl-long.451>.
- 562 H. Wimmer and J. Perner. Beliefs about beliefs: Representation and constraining function of wrong
563 beliefs in young children’s understanding of deception. *Cognition*, 13(1):103–128, 1983.

- 564 Y. Wu, Y. He, Y. Jia, R. Mihalcea, Y. Chen, and N. Deng. Hi-ToM: A benchmark for evaluating
565 higher-order theory of mind reasoning in large language models. In H. Bouamor, J. Pino, and
566 K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages
567 10691–10706, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/
568 2023.findings-emnlp.717. URL <https://aclanthology.org/2023.findings-emnlp.717>.
- 569 H. Xu, R. Zhao, L. Zhu, J. Du, and Y. He. OpenToM: A comprehensive benchmark for evaluating
570 theory-of-mind reasoning capabilities of large language models. In L.-W. Ku, A. Martins, and
571 V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational
572 Linguistics (Volume 1: Long Papers)*, pages 8593–8623, Bangkok, Thailand, Aug. 2024. Associa-
573 tion for Computational Linguistics. URL <https://aclanthology.org/2024.acl-long.466>.
- 574 P. Zhou, A. Madaan, S. P. Potharaju, A. Gupta, K. R. McKee, A. Holtzman, J. Pujara, X. Ren,
575 S. Mishra, A. Nematzadeh, et al. How far are large language models from agents with theory-of-
576 mind? *arXiv preprint arXiv:2310.03051*, 2023.
- 577 W. Zhu, Z. Zhang, and Y. Wang. Language models represent beliefs of self and others. *arXiv preprint
578 arXiv:2402.18496*, 2024.

579 References

- 580 G. Alain. Understanding intermediate layers using linear classifier probes. *arXiv preprint
581 arXiv:1610.01644*, 2016.
- 582 S. Baron-Cohen, A. M. Leslie, and U. Frith. Does the autistic child have a “theory of mind”?
583 *Cognition*, 21(1):37–46, 1985.
- 584 Y. Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*,
585 48(1):207–219, 2022.
- 586 M. Bortoletto, C. Ruhdorfer, L. Shi, and A. Bulling. Benchmarking mental state representations in
587 language models. *arXiv preprint arXiv:2406.17513*, 2024.
- 588 C. Chan, C. Jiayang, Y. Yim, Z. Deng, W. Fan, H. Li, X. Liu, H. Zhang, W. Wang, and Y. Song. Ne-
589 gotiationtom: A benchmark for stress-testing machine theory of mind on negotiation surrounding.
590 *arXiv preprint arXiv:2404.13627*, 2024.
- 591 Q. Dai, B. Heinzerling, and K. Inui. Representational analysis of binding in language models.
592 In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on
593 Empirical Methods in Natural Language Processing*, pages 17468–17493, Miami, Florida, USA,
594 Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.967.
595 URL <https://aclanthology.org/2024.emnlp-main.967/>.
- 596 X. Davies, M. Nadeau, N. Prakash, T. R. Shaham, and D. Bau. Discovering variable binding circuitry
597 with desiderata, 2023. URL <https://arxiv.org/abs/2307.03637>.
- 598 N. De Cao, M. S. Schlichtkrull, W. Aziz, and I. Titov. How do decisions emerge across layers in
599 neural models? interpretation with differentiable masking. In *Proceedings of the 2020 Conference
600 on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3243–3255, Online, Nov.
601 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.262. URL
602 <https://aclanthology.org/2020.emnlp-main.262>.
- 603 D. C. Dennett. *The Intentional Stance*. MIT Press, 1981.
- 604 N. Elhage, N. Nanda, C. Olsson, T. Henighan, N. Joseph, B. Mann, A. Askell, Y. Bai, A. Chen,
605 T. Conerly, N. DasSarma, D. Drain, D. Ganguli, Z. Hatfield-Dodds, D. Hernandez, A. Jones,
606 J. Kernion, L. Lovitt, K. Ndousse, D. Amodei, T. Brown, J. Clark, J. Kaplan, S. McCandlish, and
607 C. Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.
608 <https://transformer-circuits.pub/2021/framework/index.html>.
- 609 J. Feng and J. Steinhardt. How do language models bind entities in context? *arXiv preprint
610 arXiv:2310.17191*, 2023.

- 611 J. Feng, S. Russell, and J. Steinhardt. Monitoring latent world states in language models with
612 propositional probes. *CoRR*, abs/2406.19501, 2024. doi: 10.48550/ARXIV.2406.19501. URL
613 <https://doi.org/10.48550/arXiv.2406.19501>.
- 614 M. Finlayson, A. Mueller, S. Gehrmann, S. M. Shieber, T. Linzen, and Y. Belinkov. Causal analysis
615 of syntactic agreement mechanisms in neural language models. In C. Zong, F. Xia, W. Li, and
616 R. Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computa-*
617 *tional Linguistics and the 11th International Joint Conference on Natural Language Processing,*
618 *ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 1828–1843.
619 Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.ACL-LONG.144. URL
620 <https://doi.org/10.18653/v1/2021.acl-long.144>.
- 621 J. F. Fiotto-Kaufman, A. R. Loftus, E. Todd, J. Brinkmann, K. Pal, D. Troitskii, M. Ripa, A. Belfki,
622 C. Rager, C. Juang, A. Mueller, S. Marks, A. S. Sharma, F. Lucchetti, N. Prakash, C. E. Brodley,
623 A. Guha, J. Bell, B. C. Wallace, and D. Bau. NNSight and NDIF: Democratizing access to
624 open-weight foundation model internals. In *The Thirteenth International Conference on Learning*
625 *Representations*, 2025. URL <https://openreview.net/forum?id=MxbEiFRf39>.
- 626 K. Gandhi, J.-P. Fränken, T. Gerstenberg, and N. Goodman. Understanding social reasoning in
627 language models with language models. *Advances in Neural Information Processing Systems*, 36,
628 2024.
- 629 A. Geiger, K. Richardson, and C. Potts. Neural natural language inference models partially em-
630 bed theories of lexical entailment and negation. In A. Alishahi, Y. Belinkov, G. Chrupała,
631 D. Hupkes, Y. Pinter, and H. Sajjad, editors, *Proceedings of the Third BlackboxNLP Workshop*
632 *on Analyzing and Interpreting Neural Networks for NLP*, pages 163–173, Online, Nov. 2020.
633 Association for Computational Linguistics. doi: 10.18653/v1/2020.blackboxnlp-1.16. URL
634 <https://aclanthology.org/2020.blackboxnlp-1.16>.
- 635 A. Geiger, H. Lu, T. Icard, and C. Potts. Causal abstractions of neural networks. In
636 M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, editors,
637 *Advances in Neural Information Processing Systems 34: Annual Conference on Neu-*
638 *ral Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*,
639 pages 9574–9586, 2021. URL [https://proceedings.neurips.cc/paper/2021/hash/](https://proceedings.neurips.cc/paper/2021/hash/4f5c422f4d49a5a807eda27434231040-Abstract.html)
640 [4f5c422f4d49a5a807eda27434231040-Abstract.html](https://proceedings.neurips.cc/paper/2021/hash/4f5c422f4d49a5a807eda27434231040-Abstract.html).
- 641 A. Geiger, Z. Wu, H. Lu, J. Rozner, E. Kreiss, T. Icard, N. Goodman, and C. Potts. Inducing causal
642 structure for interpretable neural networks. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari,
643 G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine*
644 *Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 7324–7338. PMLR,
645 17–23 Jul 2022. URL <https://proceedings.mlr.press/v162/geiger22a.html>.
- 646 A. Geiger, D. Ibeling, A. Zur, M. Chaudhary, S. Chauhan, J. Huang, A. Arora, Z. Wu, N. Goodman,
647 C. Potts, and T. Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability,
648 2024. URL <https://arxiv.org/abs/2301.04709>.
- 649 M. Geva, J. Bastings, K. Filippova, and A. Globerson. Dissecting recall of factual associations in
650 auto-regressive language models, 2023. URL <https://arxiv.org/abs/2304.14767>.
- 651 A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur,
652 A. Schelten, A. Vaughan, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Sra-
653 vankumar, A. Korenev, A. Hinsvark, A. Rao, A. Zhang, A. Rodriguez, A. Gregerson, A. Spataru,
654 B. Roziere, B. Biron, B. Tang, B. Chern, C. Caucheteux, C. Nayak, C. Bi, C. Marra, C. McConnell,
655 C. Keller, C. Touret, C. Wu, C. Wong, C. C. Ferrer, C. Nikolaidis, D. Allonsius, D. Song, D. Pintz,
656 D. Livshits, D. Wyatt, D. Esiobu, D. Choudhary, D. Mahajan, D. Garcia-Olano, D. Perino, D. Hup-
657 kes, E. Lakomkin, E. AlBadawy, E. Lobanova, E. Dinan, E. M. Smith, F. Radenovic, F. Guzmán,
658 F. Zhang, G. Synnaeve, G. Lee, G. L. Anderson, G. Thattai, G. Nail, G. Mialon, G. Pang, G. Cu-
659 curell, H. Nguyen, H. Korevaar, H. Xu, H. Touvron, I. Zarov, I. A. Ibarra, I. Kloumann, I. Misra,
660 I. Evtimov, J. Zhang, J. Copet, J. Lee, J. Geffert, J. Vranes, J. Park, J. Mahadeokar, J. Shah,
661 J. van der Linde, J. Billock, J. Hong, J. Lee, J. Fu, J. Chi, J. Huang, J. Liu, J. Wang, J. Yu, J. Bitton,
662 J. Spisak, J. Park, J. Rocca, J. Johnston, J. Saxe, J. Jia, K. V. Alwala, K. Prasad, K. Upasani, K. Plaw-
663 iak, K. Li, K. Heafield, K. Stone, K. El-Arini, K. Iyer, K. Malik, K. Chiu, K. Bhalla, K. Lakhota,

664 L. Rantala-Yearly, L. van der Maaten, L. Chen, L. Tan, L. Jenkins, L. Martin, L. Madaan, L. Malo,
665 L. Blecher, L. Landzaat, L. de Oliveira, M. Muzzi, M. Pasupuleti, M. Singh, M. Paluri, M. Kar-
666 das, M. Tsimpoukelli, M. Oldham, M. Rita, M. Pavlova, M. Kambadur, M. Lewis, M. Si, M. K.
667 Singh, M. Hassan, N. Goyal, N. Torabi, N. Bashlykov, N. Bogoychev, N. Chatterji, N. Zhang,
668 O. Duchenne, O. Çelebi, P. Alrassy, P. Zhang, P. Li, P. Vasic, P. Weng, P. Bhargava, P. Dubal, P. Kr-
669 ishnan, P. S. Koura, P. Xu, Q. He, Q. Dong, R. Srinivasan, R. Ganapathy, R. Calderer, R. S. Cabral,
670 R. Stojnic, R. Raileanu, R. Maheswari, R. Girdhar, R. Patel, R. Sauvestre, R. Polidoro, R. Sumbaly,
671 R. Taylor, R. Silva, R. Hou, R. Wang, S. Hosseini, S. Chennabasappa, S. Singh, S. Bell, S. S. Kim,
672 S. Edunov, S. Nie, S. Narang, S. Raparthy, S. Shen, S. Wan, S. Bhosale, S. Zhang, S. Vandenhende,
673 S. Batra, S. Whitman, S. Sootla, S. Collot, S. Gururangan, S. Borodinsky, T. Herman, T. Fowler,
674 T. Sheasha, T. Georgiou, T. Scialom, T. Speckbacher, T. Mihaylov, T. Xiao, U. Karn, V. Goswami,
675 V. Gupta, V. Ramanathan, V. Kerkez, V. Gonguet, V. Do, V. Vogeti, V. Albiero, V. Petrovic, W. Chu,
676 W. Xiong, W. Fu, W. Meers, X. Martinet, X. Wang, X. Wang, X. E. Tan, X. Xia, X. Xie, X. Jia,
677 X. Wang, Y. Goldschlag, Y. Gaur, Y. Babaei, Y. Wen, Y. Song, Y. Zhang, Y. Li, Y. Mao, Z. D.
678 Coudert, Z. Yan, Z. Chen, Z. Papakipos, A. Singh, A. Srivastava, A. Jain, A. Kelsey, A. Shajnfeld,
679 A. Gangidi, A. Victoria, A. Goldstand, A. Menon, A. Sharma, A. Boesenberg, A. Baevski, A. Fein-
680 stein, A. Kallet, A. Sangani, A. Teo, A. Yunus, A. Lupu, A. Alvarado, A. Caples, A. Gu, A. Ho,
681 A. Poulton, A. Ryan, A. Ramchandani, A. Dong, A. Franco, A. Goyal, A. Saraf, A. Chowdhury,
682 A. Gabriel, A. Bharambe, A. Eisenman, A. Yazdan, B. James, B. Maurer, B. Leonhardi, B. Huang,
683 B. Loyd, B. D. Paola, B. Paranjape, B. Liu, B. Wu, B. Ni, B. Hancock, B. Wasti, B. Spence,
684 B. Stojkovic, B. Gamido, B. Montalvo, C. Parker, C. Burton, C. Mejia, C. Liu, C. Wang, C. Kim,
685 C. Zhou, C. Hu, C.-H. Chu, C. Cai, C. Tindal, C. Feichtenhofer, C. Gao, D. Civin, D. Beaty,
686 D. Kreymer, D. Li, D. Adkins, D. Xu, D. Testuggine, D. David, D. Parikh, D. Liskovich, D. Foss,
687 D. Wang, D. Le, D. Holland, E. Dowling, E. Jamil, E. Montgomery, E. Presani, E. Hahn, E. Wood,
688 E.-T. Le, E. Brinkman, E. Arcaute, E. Dunbar, E. Smothers, F. Sun, F. Kreuk, F. Tian, F. Kokkinos,
689 F. Ozgenel, F. Caggioni, F. Kanayet, F. Seide, G. M. Florez, G. Schwarz, G. Badeer, G. Swee,
690 G. Halpern, G. Herman, G. Sizov, Guangyi, Zhang, G. Lakshminarayanan, H. Inan, H. Shojanazeri,
691 H. Zou, H. Wang, H. Zha, H. Habeeb, H. Rudolph, H. Suk, H. Aspegren, H. Goldman, H. Zhan,
692 I. Damlaj, I. Molybog, I. Tufanov, I. Leontiadis, I.-E. Veliche, I. Gat, J. Weissman, J. Geboski,
693 J. Kohli, J. Lam, J. Asher, J.-B. Gaya, J. Marcus, J. Tang, J. Chan, J. Zhen, J. Reizenstein, J. Teboul,
694 J. Zhong, J. Jin, J. Yang, J. Cummings, J. Carvill, J. Shepard, J. McPhie, J. Torres, J. Ginsburg,
695 J. Wang, K. Wu, K. H. U, K. Saxena, K. Khandelwal, K. Zand, K. Matosich, K. Veeraraghavan,
696 K. Michelena, K. Li, K. Jagadeesh, K. Huang, K. Chawla, K. Huang, L. Chen, L. Garg, L. A.
697 L. Silva, L. Bell, L. Zhang, L. Guo, L. Yu, L. Moshkovich, L. Wehrstedt, M. Khabsa, M. Avalani,
698 M. Bhatt, M. Mankus, M. Hasson, M. Lennie, M. Reso, M. Groshev, M. Naumov, M. Lathi,
699 M. Keneally, M. Liu, M. L. Seltzer, M. Valko, M. Restrepo, M. Patel, M. Vyatskov, M. Samvelyan,
700 M. Clark, M. Macey, M. Wang, M. J. Hermoso, M. Metanat, M. Rastegari, M. Bansal, N. San-
701 thanam, N. Parks, N. White, N. Bawa, N. Singhal, N. Egebo, N. Usunier, N. Mehta, N. P. Laptev,
702 N. Dong, N. Cheng, O. Chernoguz, O. Hart, O. Salpekar, O. Kalinli, P. Kent, P. Parekh, P. Saab,
703 P. Balaji, P. Rittner, P. Bontrager, P. Roux, P. Dollar, P. Zvyagina, P. Ratanchandani, P. Yuvraj,
704 Q. Liang, R. Alao, R. Rodriguez, R. Ayub, R. Murthy, R. Nayani, R. Mitra, R. Parthasarathy,
705 R. Li, R. Hogan, R. Battey, R. Wang, R. Howes, R. Rinott, S. Mehta, S. Siby, S. J. Bondu,
706 S. Datta, S. Chugh, S. Hunt, S. Dhillon, S. Sidorov, S. Pan, S. Mahajan, S. Verma, S. Yamamoto,
707 S. Ramaswamy, S. Lindsay, S. Lindsay, S. Feng, S. Lin, S. C. Zha, S. Patil, S. Shankar, S. Zhang,
708 S. Zhang, S. Wang, S. Agarwal, S. Sajuyigbe, S. Chintala, S. Max, S. Chen, S. Kehoe, S. Satterfield,
709 S. Govindaprasad, S. Gupta, S. Deng, S. Cho, S. Virk, S. Subramanian, S. Choudhury, S. Goldman,
710 T. Remez, T. Glaser, T. Best, T. Koehler, T. Robinson, T. Li, T. Zhang, T. Matthews, T. Chou,
711 T. Shaked, V. Vontimitta, V. Ajayi, V. Montanez, V. Mohan, V. S. Kumar, V. Mangla, V. Ionescu,
712 V. Poenaru, V. T. Mihailescu, V. Ivanov, W. Li, W. Wang, W. Jiang, W. Bouaziz, W. Constable,
713 X. Tang, X. Wu, X. Wang, X. Wu, X. Gao, Y. Kleinman, Y. Chen, Y. Hu, Y. Jia, Y. Qi, Y. Li,
714 Y. Zhang, Y. Zhang, Y. Adi, Y. Nam, Yu, Wang, Y. Zhao, Y. Hao, Y. Qian, Y. Li, Y. He, Z. Rait,
715 Z. DeVito, Z. Rosnbrick, Z. Wen, Z. Yang, Z. Zhao, and Z. Ma. The llama 3 herd of models, 2024.
716 URL <https://arxiv.org/abs/2407.21783>.

717 H. Gweon, J. Fan, and B. Kim. Socially intelligent machines that learn from humans and help humans
718 learn. *Philosophical Transactions of the Royal Society A*, 381(2251):20220048, 2023.

719 D. A. Herrmann and B. A. Levinstein. Standards for belief representations in llms. *arXiv preprint*
720 *arXiv:2405.21030*, 2024.

721 G. Hou, W. Zhang, Y. Shen, L. Wu, and W. Lu. TimeToM: Temporal space is the key to unlocking
722 the door of large language models’ theory-of-mind. In L.-W. Ku, A. Martins, and V. Srikumar,
723 editors, *Findings of the Association for Computational Linguistics ACL 2024*, pages 11532–11547,
724 Bangkok, Thailand and virtual meeting, Aug. 2024. Association for Computational Linguistics.
725 URL <https://aclanthology.org/2024.findings-acl.685>.

726 J. Hu, F. Sosa, and T. Ullman. Re-evaluating theory of mind evaluation in large language models.
727 *arXiv preprint arXiv:2502.21098*, 2025.

728 C. Jin, Y. Wu, J. Cao, J. Xiang, Y.-L. Kuo, Z. Hu, T. Ullman, A. Torralba, J. Tenenbaum, and T. Shu.
729 MMTOM-QA: Multimodal theory of mind question answering. In L.-W. Ku, A. Martins, and
730 V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational
731 Linguistics (Volume 1: Long Papers)*, pages 16077–16102, Bangkok, Thailand, Aug. 2024. Associa-
732 tion for Computational Linguistics. URL <https://aclanthology.org/2024.acl-long.851>.

733 H. Kim, M. Sclar, X. Zhou, R. Bras, G. Kim, Y. Choi, and M. Sap. FANTOM: A benchmark for
734 stress-testing machine theory of mind in interactions. In H. Bouamor, J. Pino, and K. Bali, editors,
735 *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages
736 14397–14413, Singapore, Dec. 2023a. Association for Computational Linguistics. doi: 10.18653/
737 v1/2023.emnlp-main.890. URL <https://aclanthology.org/2023.emnlp-main.890/>.

738 H. Kim, M. Sclar, X. Zhou, R. L. Bras, G. Kim, Y. Choi, and M. Sap. Fantom: A benchmark for
739 stress-testing machine theory of mind in interactions. *arXiv preprint arXiv:2310.15421*, 2023b.

740 N. Kim and S. Schuster. Entity tracking in language models. *arXiv preprint arXiv:2305.02363*, 2023.

741 Y. Kim and S. S. Sundar. Anthropomorphism of computers: Is it mindful or mindless? *Computers in
742 Human Behavior*, 28(1):241–250, 2012.

743 M. Kosinski. Evaluating large language models in theory of mind tasks. *Proceedings of the National
744 Academy of Sciences*, 121(45), Oct. 2024. ISSN 1091-6490. doi: 10.1073/pnas.2405460121. URL
745 <http://dx.doi.org/10.1073/pnas.2405460121>.

746 M. Le, Y.-L. Boureau, and M. Nickel. Revisiting the evaluation of theory of mind through question
747 answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language
748 Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-
749 IJCNLP)*, pages 5872–5877, 2019.

750 B. Z. Li, M. Nye, and J. Andreas. Implicit representations of meaning in neural language models.
751 In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Proceedings of the 59th Annual Meeting
752 of the Association for Computational Linguistics and the 11th International Joint Conference
753 on Natural Language Processing (Volume 1: Long Papers)*, pages 1813–1827, Online, Aug.
754 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.143. URL
755 <https://aclanthology.org/2021.acl-long.143>.

756 K. Li, O. Patel, F. Viégas, H. Pfister, and M. Wattenberg. Inference-time intervention: Eliciting
757 truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36,
758 2024.

759 T. Lieberum, M. Rahtz, J. Kramár, N. Nanda, G. Irving, R. Shah, and V. Mikulik. Does circuit
760 analysis interpretability scale? evidence from multiple choice capabilities in chinchilla, 2023. URL
761 <https://arxiv.org/abs/2307.09458>.

762 K. Meng, D. Bau, A. Andonian, and Y. Belinkov. Locating and editing factual associations in GPT.
763 *Advances in Neural Information Processing Systems*, 36, 2022.

764 S. R. Moghaddam and C. J. Honey. Boosting theory-of-mind performance in large language models
765 via prompting. *arXiv preprint arXiv:2304.11490*, 2023.

766 A. Mueller, J. Brinkmann, M. Li, S. Marks, K. Pal, N. Prakash, C. Rager, A. Sankaranarayanan, A. S.
767 Sharma, J. Sun, E. Todd, D. Bau, and Y. Belinkov. The quest for the right mediator: A history,
768 survey, and theoretical grounding of causal interpretability, 2024. URL [https://arxiv.org/
769 abs/2408.01416](https://arxiv.org/abs/2408.01416).

- 770 C. Olsson, N. Elhage, N. Nanda, N. Joseph, N. DasSarma, T. Henighan, B. Mann, A. Askell, Y. Bai,
771 A. Chen, T. Conerly, D. Drain, D. Ganguli, Z. Hatfield-Dodds, D. Hernandez, S. Johnston, A. Jones,
772 J. Kernion, L. Lovitt, K. Ndousse, D. Amodei, T. Brown, J. Clark, J. Kaplan, S. McCandlish,
773 and C. Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022.
774 <https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html>.
- 775 N. Prakash, T. R. Shaham, T. Haklay, Y. Belinkov, and D. Bau. Fine-tuning enhances existing
776 mechanisms: A case study on entity tracking. In *Proceedings of the 2024 International Conference*
777 *on Learning Representations*, 2024. arXiv:2402.14811.
- 778 D. Premack and G. Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and Brain*
779 *Sciences*, 1(4):515–526, 1978. doi: 10.1017/S0140525X00076512.
- 780 C. Riedl, Y. J. Kim, P. Gupta, T. W. Malone, and A. W. Woolley. Quantifying collective intelligence
781 in human groups. *Proceedings of the National Academy of Sciences*, 118(21):e2005737118, 2021.
- 782 N. Rimskey, N. Gabrieli, J. Schulz, M. Tong, E. Hubinger, and A. M. Turner. Steering llama 2 via
783 contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2023.
- 784 N. Saphra and S. Wiegrefe. Mechanistic? In Y. Belinkov, N. Kim, J. Jumelet, H. Mohebbi,
785 A. Mueller, and H. Chen, editors, *Proceedings of the 7th BlackboxNLP Workshop: Analyzing*
786 *and Interpreting Neural Networks for NLP*, pages 480–498, Miami, Florida, US, Nov. 2024.
787 Association for Computational Linguistics. doi: 10.18653/v1/2024.blackboxnlp-1.30. URL
788 <https://aclanthology.org/2024.blackboxnlp-1.30/>.
- 789 M. Sclar, S. Kumar, P. West, A. Suhr, Y. Choi, and Y. Tsvetkov. Minding language models’(lack of)
790 theory of mind: A plug-and-play multi-character belief tracker. *arXiv preprint arXiv:2306.00924*,
791 2023.
- 792 M. Sclar, J. Yu, M. Fazel-Zarandi, Y. Tsvetkov, Y. Bisk, Y. Choi, and A. Celikyilmaz. Explore theory
793 of mind: Program-guided adversarial data generation for theory of mind reasoning. *ICLR*, 2025.
- 794 N. Shapira, G. Zwirn, and Y. Goldberg. How well do large language models perform on faux
795 pas tests? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages
796 10438–10451, Toronto, Canada, July 2023. Association for Computational Linguistics. URL
797 <https://aclanthology.org/2023.findings-acl.663>.
- 798 N. Shapira, M. Levy, S. H. Alavi, X. Zhou, Y. Choi, Y. Goldberg, M. Sap, and V. Shwartz. Clever hans
799 or neural theory of mind? stress testing social reasoning in large language models. In Y. Graham and
800 M. Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association*
801 *for Computational Linguistics (Volume 1: Long Papers)*, pages 2257–2273, St. Julian’s, Malta,
802 Mar. 2024. Association for Computational Linguistics. URL [https://aclanthology.org/](https://aclanthology.org/2024.eacl-long.138)
803 [2024.eacl-long.138](https://aclanthology.org/2024.eacl-long.138).
- 804 P. Smolensky. Neural and conceptual interpretation of PDP models. In J. L. McClelland, D. E.
805 Rumelhart, and the PDP Research Group, editors, *Parallel Distributed Processing: Explorations in*
806 *the Microstructure of Cognition: Psychological and Biological Models*, volume 2, pages 390–431.
807 MIT Press, 1986.
- 808 J. W. Strachan, D. Albergo, G. Borghini, O. Pansardi, E. Scaliti, S. Gupta, K. Saxena, A. Rufo,
809 S. Panzeri, G. Manzi, et al. Testing theory of mind in large language models and humans. *Nature*
810 *Human Behaviour*, 8(7):1285–1295, 2024a.
- 811 J. W. A. Strachan, D. Albergo, G. Borghini, O. Pansardi, E. Scaliti, S. Gupta, K. Saxena, A. Rufo,
812 S. Panzeri, G. Manzi, M. S. A. Graziano, and C. Becchio. Testing theory of mind in large language
813 models and humans. *Nature Human Behaviour*, 8(7):1285–1295, Jul 2024b. ISSN 2397-3374. doi:
814 [10.1038/s41562-024-01882-z](https://doi.org/10.1038/s41562-024-01882-z). URL <https://doi.org/10.1038/s41562-024-01882-z>.
- 815 W. Street, J. O. Siy, G. Keeling, A. Baranes, B. Barnett, M. McKibben, T. Kanyere, A. Lentz, B. A.
816 y Arcas, and R. I. M. Dunbar. Llms achieve adult human performance on higher-order theory of
817 mind tasks, 2024. URL <https://arxiv.org/abs/2405.18870>.

818 J. Vig, S. Gehrmann, Y. Belinkov, S. Qian, D. Nevo, Y. Singer, and S. Shieber. Investigating gender
819 bias in language models using causal mediation analysis. In H. Larochelle, M. Ranzato, R. Hadsell,
820 M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33,
821 pages 12388–12401. Curran Associates, Inc., 2020. URL [https://proceedings.neurips.cc/
822 paper_files/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf).

823 K. R. Wang, A. Variengien, A. Conmy, B. Shlegeris, and J. Steinhardt. Interpretability in the wild: a
824 circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference
825 on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.
826 URL <https://openreview.net/forum?id=NpsVSN6o4ul>.

827 A. Wilf, S. Lee, P. P. Liang, and L.-P. Morency. Think twice: Perspective-taking improves large
828 language models’ theory-of-mind capabilities. In L.-W. Ku, A. Martins, and V. Srikumar, editors,
829 *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume
830 1: Long Papers)*, pages 8292–8308, Bangkok, Thailand, Aug. 2024. Association for Computational
831 Linguistics. URL <https://aclanthology.org/2024.acl-long.451>.

832 H. Wimmer and J. Perner. Beliefs about beliefs: Representation and constraining function of wrong
833 beliefs in young children’s understanding of deception. *Cognition*, 13(1):103–128, 1983.

834 Y. Wu, Y. He, Y. Jia, R. Mihalcea, Y. Chen, and N. Deng. Hi-ToM: A benchmark for evaluating
835 higher-order theory of mind reasoning in large language models. In H. Bouamor, J. Pino, and
836 K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages
837 10691–10706, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/
838 2023.findings-emnlp.717. URL <https://aclanthology.org/2023.findings-emnlp.717>.

839 H. Xu, R. Zhao, L. Zhu, J. Du, and Y. He. OpenToM: A comprehensive benchmark for evaluating
840 theory-of-mind reasoning capabilities of large language models. In L.-W. Ku, A. Martins, and
841 V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational
842 Linguistics (Volume 1: Long Papers)*, pages 8593–8623, Bangkok, Thailand, Aug. 2024. Associa-
843 tion for Computational Linguistics. URL <https://aclanthology.org/2024.acl-long.466>.

844 P. Zhou, A. Madaan, S. P. Potharaju, A. Gupta, K. R. McKee, A. Holtzman, J. Pujara, X. Ren,
845 S. Mishra, A. Nematzadeh, et al. How far are large language models from agents with theory-of-
846 mind? *arXiv preprint arXiv:2310.03051*, 2023.

847 W. Zhu, Z. Zhang, and Y. Wang. Language models represent beliefs of self and others. *arXiv preprint
848 arXiv:2402.18496*, 2024.

849 A Full prompt

No Visibility

Instruction: 1. Track the belief of each character as described in the story. 2. A character’s belief is formed only when they perform an action themselves or can observe the action taking place. 3. A character does not have any beliefs about the container and its contents which they cannot observe. 4. To answer the question, predict only what is inside the queried container, strictly based on the belief of the character, mentioned in the question. 5. If the queried character has no belief about the container in question, then predict ‘unknown’. 6. Do not predict container or character as the final output.

Story: 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 Bob believe the bottle contains?

Answer:

850

Explicit Visibility

Instruction: 1. Track the belief of each character as described in the story. 2. A character’s belief is formed only when they perform an action themselves or can observe the action taking place. 3. A character does not have any beliefs about the container and its contents which they cannot observe. 4. To answer the question, predict only what is inside the queried container, strictly based on the belief of the character, mentioned in the question. 5. If the queried character has no belief about the container in question, then predict ‘unknown’. 6. Do not predict container or character as the final output.

Story: 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. Bob can observe Carla’s actions. Carla cannot observe Bob’s actions.

Question: What does Bob believe the cup contains?

Answer:

851

852 B Related Work

853 **Theory of mind in LMs** A large body of work has focused on benchmarking different aspects of
 854 ToM through various tasks that attempt to assess LMs’ performance such as Le et al. [2019], Xu
 855 et al. [2024], Shapira et al. [2023], Jin et al. [2024], Wu et al. [2023], Kim et al. [2023b], Chan et al.
 856 [2024], Strachan et al. [2024a] and many more. In addition, there are various methods tailored to
 857 improve ToM ability in LMs through prompting [e.g., Sclar et al., 2023, Zhou et al., 2023, Wilf et al.,
 858 2024, Moghaddam and Honey, 2023, Hou et al., 2024].

859 **Entity tracking in LMs** Entity tracking and variable binding are crucial abilities for LMs to exhibit
 860 not only coherent ToM capabilities, but also neurosymbolic reasoning. Many existing works have
 861 attempted to decipher this ability in LMs [Li et al., 2021, Davies et al., 2023, Kim and Schuster, 2023,
 862 Prakash et al., 2024, Feng and Steinhart, 2023, Feng et al., 2024, Dai et al., 2024]. Our work builds
 863 on their empirical insights and extends the current understanding of how LMs bind various entities
 864 defined in context.

865 **Mechanistic interpretability of theory of mind** Only a few empirical studies explored the
 866 underlying mechanisms of ToM of LM [Zhu et al., 2024, Bortoletto et al., 2024] [Herrmann and
 867 Levinstein, 2024, is a notable theoretical paper]. Those studies focus on probing techniques [Belinkov,
 868 2022, Alain, 2016] to identify internal representations of beliefs and used steering techniques [Li
 869 et al., 2024, Rinsky et al., 2023] to improve LM performance by manipulating their activations.
 870 However, the mechanism by which LMs solve those tasks remains a black box, limiting our ability to
 871 understand, predict, and control LMs’ behaviors.

872 C The CausalToM Dataset

873 In total, there are 4 templates (one without and 3 with explicit visibility statements). Each template
 874 allows 4 different types of questions (CharacterX asked about ObjectY). We used lists of 103
 875 characters, 21 objects, and 23 states. In our interchange intervention experiments, we randomly
 876 sample 80 pairs of original and counterfactual stories.

877 D Causal Mediation Analysis

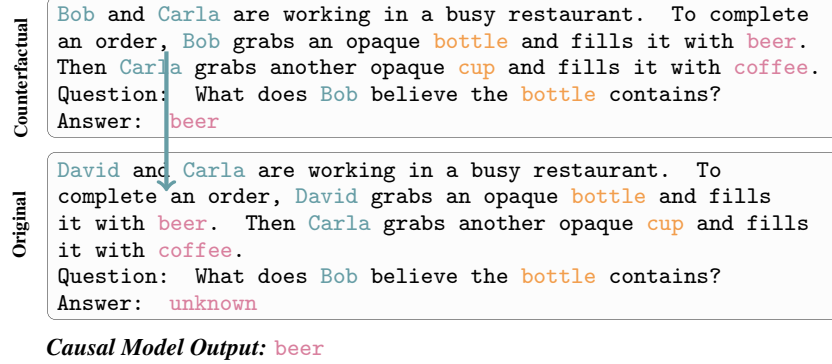


Figure 8: **Causal Mediation Analysis:** The original example produces the output *unknown* because *Bob* is not mentioned in the story, leaving the model without any information about his beliefs. However, when the residual stream vectors corresponding to *Bob* from the counterfactual run are patched into the original run, the model acquires the necessary information about that character and consequently updates its output to *beer*.

878 In addition to the experiment shown in Fig.8, we conduct similar experiments for the object and
 879 state tokens by replacing them in the story with random tokens, which alters the original example’s
 880 final output. However, patching the residual stream vectors of these tokens from the counterfactual
 881 run restores the relevant information, enabling the model to predict the causal model output. The
 882 results of these experiments are collectively presented in Fig.3, with separate heatmaps shown in
 883 Fig. 9, 10, 11.

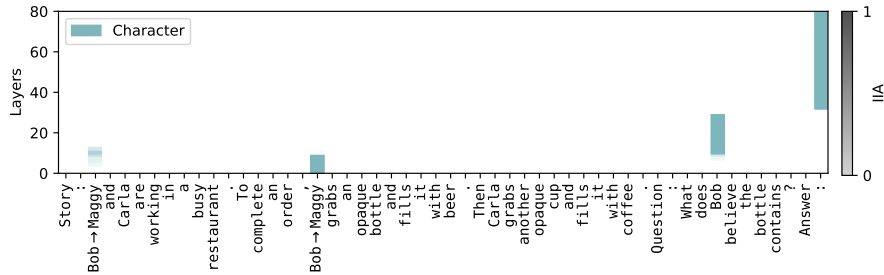


Figure 9: Information flow of character input tokens using causal mediation analysis.

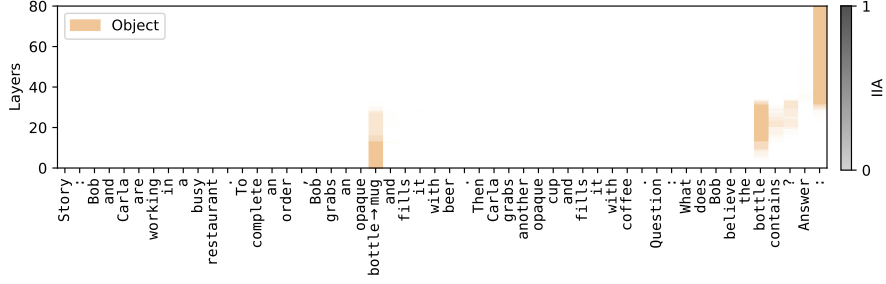


Figure 10: Information flow of object input tokens using causal mediation analysis.

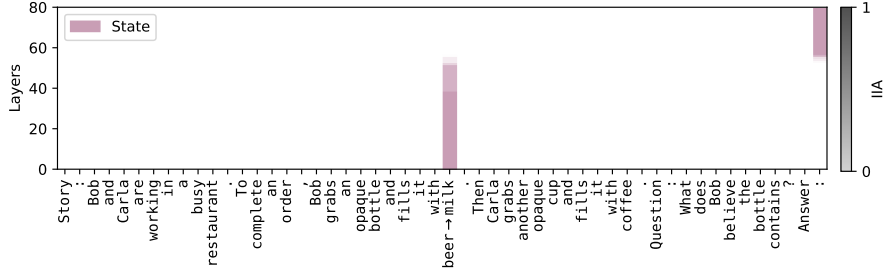


Figure 11: Information flow of state input tokens using causal mediation analysis.

884 E Desiderate Based Patching Via Causal Abstraction

885 **Causal Models and Interventions** A deterministic causal model \mathcal{M} has *variables* that take on
 886 *values*. Each variable has a *mechanism* that determines the value of the variable based on the values of
 887 *parent variables*. Variables without parents, denoted \mathbf{X} , can be thought of as inputs that determine the
 888 setting of all other variables, denoted $\mathcal{M}(\mathbf{x})$. A *hard intervention* $A \leftarrow a$ overrides the mechanisms
 889 of variable A , fixing it to a constant value a .

890 **Interchange Interventions** We perform *interchange interventions* [Vig et al., 2020, Geiger et al.,
 891 2020] where a variable (or set of features) A is fixed to be the value it would take on if the LM were
 892 processing *counterfactual input* \mathbf{c} . We write $A \leftarrow \text{Get}(\mathcal{M}(\mathbf{c}), A)$ where $\text{Get}(\mathcal{M}(\mathbf{c}), A)$ is the value
 893 of variable A when \mathcal{M} processes input \mathbf{c} . In experiments, we will feed a *original input* \mathbf{o} to a model
 894 under an interchange intervention $\mathcal{M}_{A \leftarrow \text{Get}(\mathcal{M}(\mathbf{c}), A)}(\mathbf{o})$.

895 **Featurizing Hidden Vectors** The dimensions of hidden vectors are not an ideal unit of analysis
 896 [Smolensky, 1986], and so it is typical to *featurize* a hidden vector using some invertible function,
 897 e.g., an orthogonal matrix, to project a hidden vector into a new variable space with more inter-
 898 pretable dimensions called “features” [Mueller et al., 2024]. A feature intervention $\mathbf{F}_h \leftarrow \mathbf{f}$ edits the
 899 mechanism of a hidden vector \mathbf{h} to fix the value of features \mathbf{F}_h to \mathbf{f} .

900 **Alignment** The LM is a *low-level causal model* \mathcal{L} where variables are dimensions of hidden vectors
 901 and the hypothesis about LM structure is a *high-level causal model* \mathcal{H} . An *alignment* Π assigns each
 902 high-level variable A to features of a hidden vector \mathbf{F}_h^A , e.g., orthogonal directions in the activation
 903 space of \mathbf{h} . To evaluate an alignment, we perform intervention experiments to evaluate whether
 904 high-level interventions on the variables in \mathcal{H} have the same effect as interventions on the aligned
 905 features in \mathcal{L} .

906 **Causal Abstraction** We use interchange interventions to reveal whether the hypothesized causal
 907 model \mathcal{H} is an abstraction of an LM \mathcal{L} . To simplify, assume both models share an input and output
 908 space. The high-level model \mathcal{H} is an abstraction of the low-level model \mathcal{L} under a given alignment
 909 when each high-level interchange intervention and the aligned low-level intervention result in the same
 910 output. For a high-level intervention on A aligned with low-level features \mathbf{F}_h^A with a counterfactual
 911 input \mathbf{c} and original input \mathbf{b} , we write

$$\text{GetOutput}(\mathcal{L}_{\mathbf{F}_h^A \leftarrow \text{Get}(\mathcal{L}(\mathbf{c}), \mathbf{F}_h^A)}(\mathbf{o})) = \text{GetOutput}(\mathcal{H}_{A \leftarrow \text{Get}(\mathcal{H}(\mathbf{c}), A)}(\mathbf{o})) \quad (1)$$

912 If the low-level interchange intervention on the LM produces the same output as the aligned high-level
 913 intervention on the algorithm, this is a piece of evidence in favor of the hypothesis. This extends
 914 naturally to multi-variable interventions [Geiger et al., 2024].

915 **Graded Faithfulness Metric** We construct *counterfactual datasets* for each causal variable where
 916 an example consists of a base prompt and a counterfactual prompt . The *counterfactual label* is the
 917 expected output of the algorithm after the high-level interchange intervention, i.e., the right-side of
 918 Equation 1. The interchange intervention accuracy is the proportion of examples for which Equation 1
 919 holds, i.e., the degree to which \mathcal{H} faithfully abstracts \mathcal{L} .

920 **Aligning Features to Causal Variables** In our experiments, we use Singular Vector Decomposition
 921 (SVD) to featurize residual stream vectors, i.e., features are the orthogonal singular vectors. For
 922 a given transformer layer and token location, we collect the residual stream vectors across a large
 923 number of examples and compute the singular vectors. Given singular vector features \mathbf{F}_h of a hidden
 924 vector \mathbf{h} in the residual stream of the LM \mathcal{L} , we select features to align with a causal variable A in
 925 causal model \mathcal{H} using Desiderata-based Component Masking (DCM) [De Cao et al., 2020, Davies
 926 et al., 2023, Prakash et al., 2024]. Given original input \mathbf{o} and counterfactual input \mathbf{c} , we train a mask
 927 $\mathbf{m} \in [0, 1]^{|\mathbf{F}_h|}$ on the following objective

$$\text{CE}\left(\text{GetLogits}\left(\mathcal{L}_{\mathbf{F}_h \leftarrow \mathbf{m} \circ \text{Get}(\mathcal{L}(\mathbf{c}), \mathbf{F}_h)}(\mathbf{b})\right), \text{GetLogits}\left(\mathcal{H}_{A \leftarrow \text{Get}(\mathcal{H}(\mathbf{c}), A)}(\mathbf{b})\right)\right) \quad (2)$$

Algorithm 2 High-level causal model for the no visibility

```

1: procedure BELIEFTRACKING( $c_1, o_1, s_1, c_2, o_2, s_2, q_c, q_o$ )
2:   Ordering ID assignment
3:    $c_1^{OI}, o_1^{OI}, s_1^{OI} \leftarrow \text{AssignOIs}([c_1, o_1, s_1], 1)$ 
4:    $c_2^{OI}, o_2^{OI}, s_2^{OI} \leftarrow \text{AssignOIs}([c_2, o_2, s_2], 2)$ 
5:
6:   Binding lookback mechanism
7:    $\text{binding\_address}_1 \leftarrow (\text{copy}(c_1^{OI}), \text{copy}(o_1^{OI}))$ 
8:    $\text{binding\_address}_2 \leftarrow (\text{copy}(c_2^{OI}), \text{copy}(o_2^{OI}))$ 
9:
10:   $q_c^{OI} \leftarrow \text{copy}(\{c_1 : c_1^{OI}, c_2 : c_2^{OI}\}[q_c])$ 
11:   $q_o^{OI} \leftarrow \text{copy}(\{o_1 : o_1^{OI}, o_2 : o_2^{OI}\}[q_o])$ 
12:   $\text{binding\_pointer} \leftarrow (q_c^{OI}, q_o^{OI})$ 
13:
14:  if  $\text{binding\_address}_1 = \text{binding\_pointer}$  then
15:     $\text{binding\_payload} \leftarrow \text{copy}(s_1^{OI})$ 
16:  else if  $\text{binding\_address}_2 = \text{binding\_pointer}$  then
17:     $\text{binding\_payload} \leftarrow \text{copy}(s_2^{OI})$ 
18:  end if
19:
20:  Answer lookback mechanism
21:   $\text{answer\_pointer} \leftarrow \text{binding\_payload}$ 
22:   $\text{answer1\_address} \leftarrow s_1^{OI}$ 
23:   $\text{answer2\_address} \leftarrow s_2^{OI}$ 
24:  if  $\text{answer1\_address} = \text{answer\_pointer}$  then
25:     $\text{answer\_payload} \leftarrow s_1$ 
26:  else if  $\text{answer2\_address} = \text{answer\_pointer}$  then
27:     $\text{answer\_payload} \leftarrow s_2$ 
28:  end if
29:  return  $\text{answer\_payload}$ 
30: end procedure

```

929 **G Desiderata-based Component Masking**

930 While interchange interventions on residual vectors reveal where a causal variable might be encoded
931 in the LM’s internal activations, they do not localize the variable to specific subspaces. To address
932 this, we apply the *Desiderata-based Component Masking* technique [De Cao et al., 2020, Davies et al.,
933 2023, Prakash et al., 2024], which learns a sparse binary mask \mathbf{m} over the singular vectors of the LM’s
934 internal activations. We first cache the internal activations from 500 samples at the token positions
935 specified in the main text for each experiment. Next, we apply *Singular Value Decomposition* to
936 compute the singular vectors as a matrix $V \in \mathbb{R}^{d \times 500}$ where d is the dimensionality of the residual
937 stream. We then masked this matrix using a learnable binary vector $\mathbf{m} \in [0, 1]^d$ to choose a subset of
938 singular vectors

$$V_{\text{masked}} = V\mathbf{m} \quad (3)$$

939 The chosen subset of vectors is used to construct a *projection matrix* $W_{\text{proj}} \in \mathbb{R}^{d \times d}$.

$$W_{\text{proj}} = V_{\text{masked}} V_{\text{masked}}^T \quad (4)$$

940 Then, we perform subspace-level interchange interventions (rather than replacing the entire residual
941 vector) using the following equations:

$$h_{\text{new}} = W_{\text{proj}} h_c + (I - W_{\text{proj}}) h_o \quad (5)$$

942 where h_o is the full residual stream of the original run, h_c is the full residual stream of the counterfac-
943 tual run, and h_{new} is the intervened vector where the chosen subspace of h_o is replaced with that of
944 h_c .

The core idea is to first remove the existing information from the subspace defined by the projection matrix and then insert the counterfactual information into that same subspace using the same projection matrix.

In order to find the optimal subspace, we optimize \mathbf{m} to maximize the agreement between the causal model output and the LM’s output. To do so, we train the mask for each experiment on 80 examples of the same counterfactual datasets specified in the main text and use another 80 samples as the validation set. We use the following objective function, which maximizes the logit of the causal model output token:

$$\mathcal{L} = -\text{logit}_{\text{causal_model_output_under_intervention}} + \lambda \sum \mathbf{m} \quad (6)$$

Where λ is a hyperparameter used to control the rank of the subspace and \mathbf{m} is the learnable mask. See Appendix E for details on how the causal model output under intervention are computed. We trained \mathbf{m} for one epoch with ADAM optimizer, on batches of size 4 and a learning rate of 0.01. During training, the parameters of \mathbf{m} are continuous and constrained to lie within the range $[0, 1]$. To enforce this constraint, we clamp their values after each gradient update. During evaluation, we binarize the mask by rounding each parameter to the nearest integer, i.e., 0 or 1.

H Aligning Character and Object OIs

As mentioned in section 4.2, the source information, consisting of character and object OI, is duplicated to form the address and pointer of the binding lookback. Here, we describe another experiment to verify that the source information is copied to both the address and the pointer. More specifically, we conduct the same interchange intervention experiment as described in Fig. 5, but without freezing the residual vectors at the state tokens. Based on our hypothesis, this intervention will not be able to change the state of the original run, since the intervention at the source information will affect both address and pointer, hence making the model form the original QK-circuit.

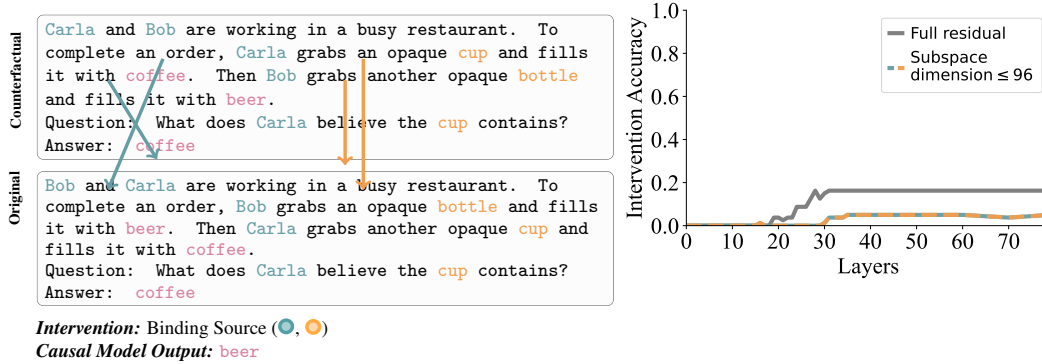


Figure 12: **Source Information** of Binding lookback: In this interchange intervention experiment, the source information—i.e., the character and object OIDs (●, ●)—is modified, while the address and payload (●, ●, ▲) are recomputed based on the modified source. Since both the address and pointer information are derived from the altered source, the binding lookback ultimately retrieves the same original state token as the payload. As a result, we do not observe high intervention accuracy.

In section 4.2, we identified the source of the information but did not fully determine the locations of each character and object OI. To address this, we now localize the character and object OIs separately to gain a clearer understanding of the layers at which they appear in the residual streams of their respective tokens, as shown in Fig.13 and Fig.14.

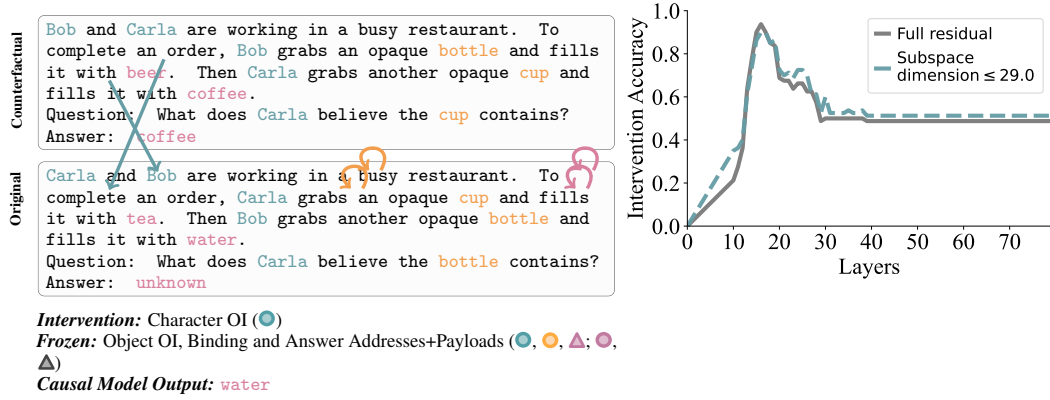


Figure 13: **Character OI**: This interchange intervention experiment swaps the character OI (●), while freezing the object OI as well as binding lookback address and payload (●, ●, ▲, ▲). Swapping the character OIs in the story tokens changes the queried character OI to the other one. Hence, the final output changes from *unknown* to *water*.

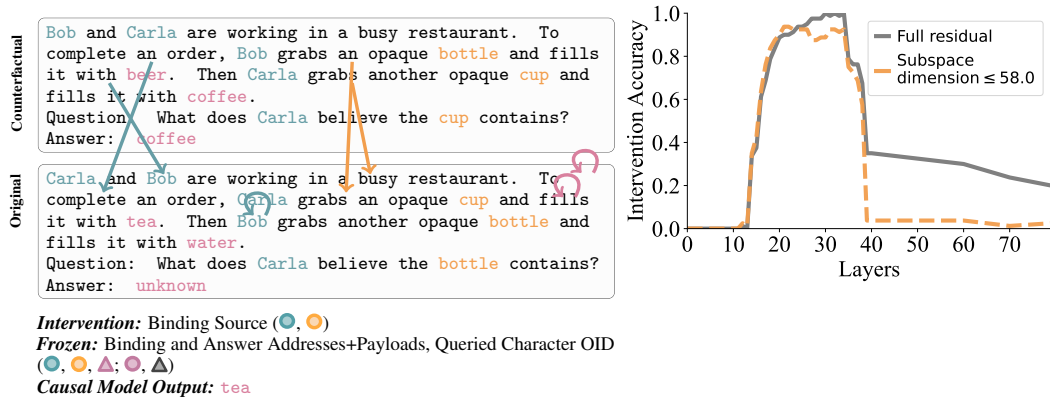


Figure 14: **Object OI**: This interchange intervention experiment swaps both the character and object OIs (●, ●), while freezing the address and payload of binding lookback (●, ●, ▲, ▲) as well as queried character OI (●). Swapping both character and object OIs in the story tokens ensures that the queried object gets the other OI. Hence, the final output changes from *unknown* to *tea*.

I Aligning Query Character and Object OIs

In section 4.2, we localized the pointer information of binding lookback. However, we found that this information is transferred to the lookback token (last token) through two intermediate tokens: the queried character and the queried object. In this section, we separately localize the OIs of the queried character and queried object, as shown in Fig. 15 and Fig. 16.

J Speculated Payload in Visibility Lookback

As mentioned in section 5, the payload of the Visibility lookback remains undetermined. In this section, we attempt to disambiguate its semantics using the Attention Knockout technique introduced in [Geva et al., 2023], which helps reveal the flow of crucial information. We apply this technique to understand which previous tokens are vital for the formation of the payload information. Specifically, we "knock out" all attention heads at all layers of the second visibility sentence, preventing them from attending to one or more of the previous sentences. Then, we allow the attention heads to attend to the knocked-out sentence one layer at a time.

If the LM is fetching vital information from the knocked-out sentence, the interchange intervention accuracy (IIA) post-knockout will decrease. Therefore, a decrease in IIA will indicate which attention

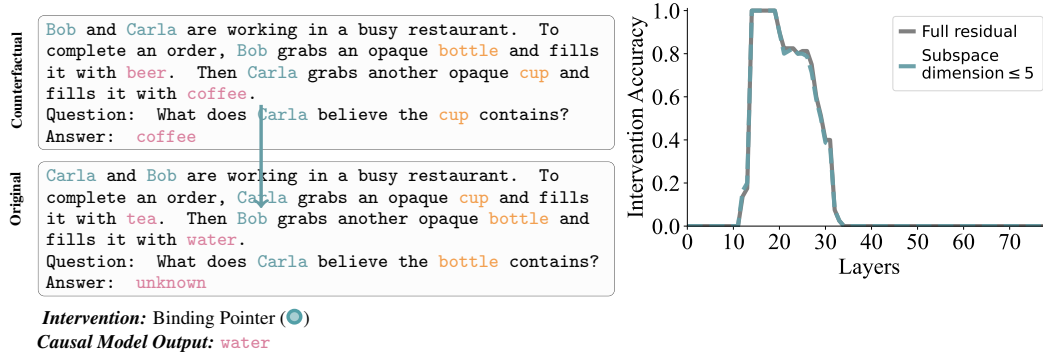


Figure 15: **Query Character OI**: This interchange intervention experiment alters the OI of the queried character (●) to the other one. Hence, the final output changes from *unknown* to *water*.

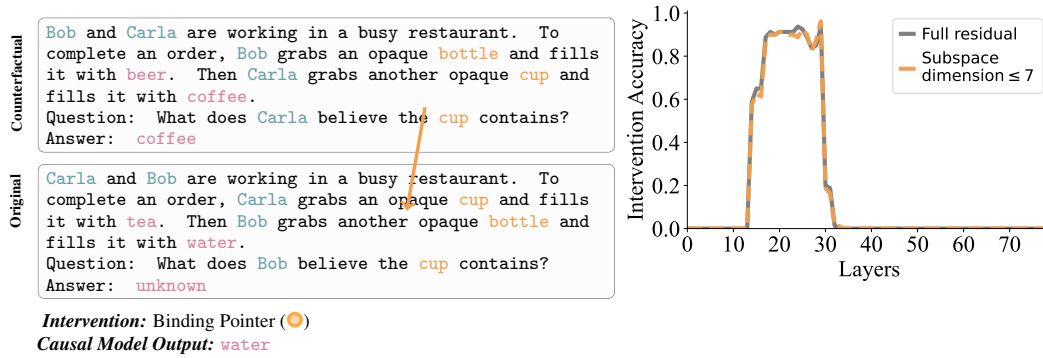


Figure 16: **Query Object OI**: This interchange intervention experiment alters the OI of the queried object (●) to the other one. Hence, the final output changes from *unknown* to *water*.

heads, at which layers, are bringing in the vital information from the knocked-out sentence. If, however, the model is not fetching any critical information from the knocked-out sentence, then knocking it out should not affect the IIA.

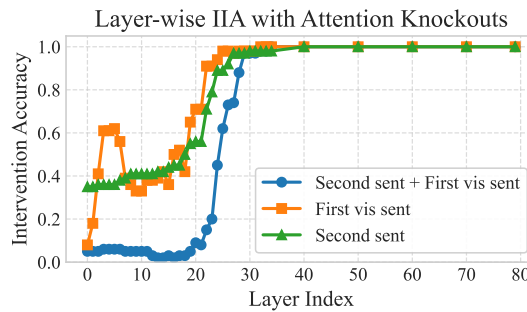


Figure 17: At the second visibility sentence, attention heads are restricted to retrieve information from one of three prior contexts: (1) both the second story sentence and the first visibility sentence (— line), (2) only the first visibility sentence (— line), or (3) only the second story sentence (— line).

To determine if any vital information is influencing the formation of the Visibility lookback payload, we perform three knockout experiments: 1) Knockout attention heads from the second visibility sentence to both the first visibility sentence and the second story sentence (which contains information about the observed character), 2) Knockout attention heads from the second visibility sentence to only the first visibility sentence, and 3) Knockout attention heads from the second visibility sentence to the second story sentence. In each experiment, we measure the effect of the knockout using IIA.

Fig.17 shows the experimental results. Knocking out any of the previous sentences affects the model’s ability to produce the correct output. The decrease in IIA in the early layers can be explained by the restriction on the movement of character OIs. Specifically, the second visibility sentence mentions the first and second characters, whose character OIs must be fetched before the model can perform any further operations. Therefore, we believe the decrease in IIA until layer 15, when the character OIs are formed (based on the results from Section H), can be attributed to the model being restricted from fetching the character OIs. However, the persistently low IIA even after this layer—especially when both the second and first visibility sentences are involved—indicates that some vital information is being fetched by the second visibility sentence, which is essential for forming the coherent Visibility lookback payload. Thus, we speculate that the Visibility payload encodes information about the observed character, specifically their character OI, which is later used to fetch the correct state OI.

K 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:

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

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:

$$Q_{\text{sv}} = \tilde{Q} \cdot S \quad (8)$$

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:

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

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

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

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. 18, 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. 19, indicating that attention heads at later layers primarily align with causal subspace containing the answer token.

L Belief Tracking Mechanism in BigToM Benchmark

This section presents preliminary evidence that the mechanisms outlined in Sections 4 and 5 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

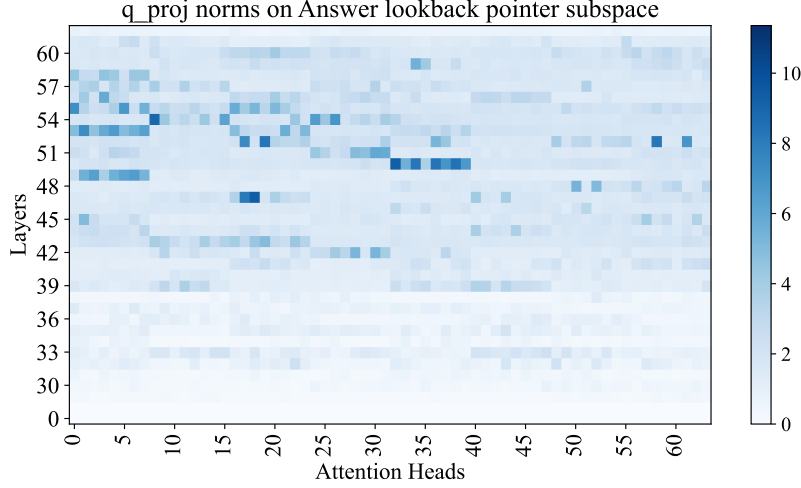


Figure 18: Alignment between the Answer lookback pointer causal subspace and the query projection matrix in Llama-3-70B-Instruct.

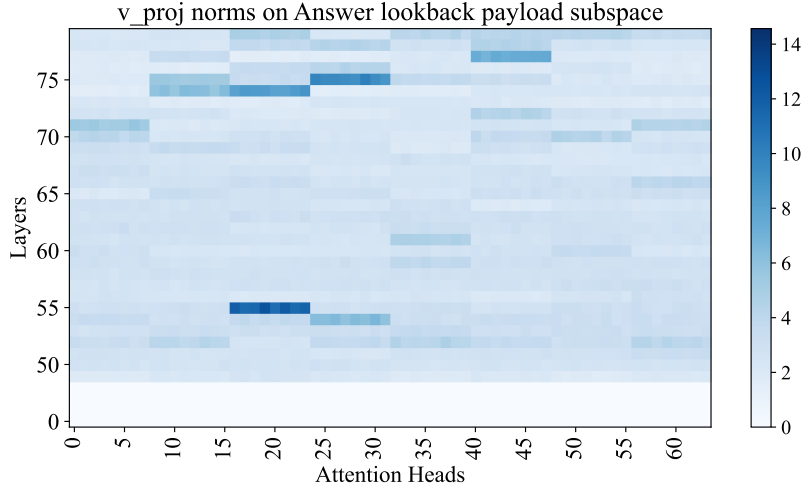


Figure 19: Alignment between the Answer lookback payload causal subspace and the value projection matrix in Llama-3-70B-Instruct.

1032 similar to that observed for CausalToM: by first converting token values to their corresponding OIs
 1033 and then performing logical operations on them using lookbacks. However, as noted in Section 3,
 1034 BigToM—like other benchmarks—lacks the coherent structure necessary for causal analysis. As
 1035 a result, we were unable to replicate all experiments conducted on CausalToM. Thus, the results
 1036 reported here provide only preliminary evidence of a similar underlying mechanism.

1037 To justify the presence of OIs, we conduct an interchange intervention experiment, similar to
 1038 the one described in Section I, aiming to localize the character OI at the character token in the
 1039 question sentence. We construct an original sample by replacing its question sentence with that of a
 1040 counterfactual sample, selected directly from the unaltered BigToM dataset. Consequently, when
 1041 processing the original sample, the model has no information about the queried character and, as
 1042 a result, produces unknown as the final output. However, if we replace the residual vector at the
 1043 queried character token in the original sample with the corresponding vector from the counterfactual
 1044 sample (which contains the character OI), the model’s output changes from unknown to the state
 1045 token(s) associated with the queried object. This is because inserting the character OI at the queried
 1046 token provides the correct pointer information, aligning with the address information at the correct
 1047 state token(s), thereby enabling the model to form the appropriate QK-circuit and retrieve the state’s
 1048 OI. As shown in Fig. 20, we observe a high IIA between layers 9 – 28—similar to the pattern seen



Figure 20: **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.

1049 in CausalToM—suggesting that the queried character token encodes the character OI in its residual
 1050 vector within these layers.

1051 Next, we investigate the Answer lookback mechanism in BigToM, focusing specifically on localizing
 1052 the pointer and payload information at the final token position. To localize the pointer information,
 1053 which encodes the correct state OI, we construct original and counterfactual samples by selecting two
 1054 completely different examples from the BigToM dataset, each with different ordered states as the
 1055 correct answer. For example, as illustrated in Fig.21, the counterfactual sample designates the first
 1056 state as the answer, **thrilling plot**, whereas the original sample designates the second state, **almond**
 1057 **milk**. We perform an intervention by swapping the residual vector at the last token position from the
 1058 counterfactual sample into the original run. The causal model outcome of this intervention is that the
 1059 model will output the alternative state token from the original sample, **oat milk**. As shown in Fig.21,
 1060 this alignment occurs between layers 33 and 51, similar to the layer range observed for the pointer
 1061 information in the Answer lookback of CausalToM.

1062 Further, to localize the payload of the Answer lookback in BigToM, we perform an interchange
 1063 intervention experiment using the same original and counterfactual samples as mentioned in the
 1064 previous experiment, but with a different expected output—namely, the correct state from the
 1065 counterfactual sample instead of the other state from the original sample. As shown in Fig. 22,
 1066 alignment emerges after layer 59, consistent with the layer range observed for the Answer lookback
 1067 payload in CausalToM.

1068 Finally, we investigate the impact of the visibility condition on the underlying mechanism and
 1069 find that, similar to CausalToM, the model uses the Visibility lookback to enhance the observing
 1070 character’s awareness based on the observed character’s actions. To localize the effect of the visibility
 1071 condition, we perform an interchange intervention in which the original and counterfactual samples
 1072 differ in belief type—that is, if the original sample involves a false belief, the counterfactual involves
 1073 a true belief, and vice versa. The expected output of this experiment is the other (incorrect) state of
 1074 the original sample. Following the methodology in Section 5, we conduct three types of interventions:
 1075 (1) only at the visibility condition sentence, (2) only at the subsequent question sentence, and (3) at
 1076 both the visibility condition and the question sentence. As shown in Fig. 23, intervening only at the
 1077 visibility sentence results in alignment at early layers, up to layer 17, while intervening only at the
 1078 subsequent question sentence leads to alignment after layer 26. Intervening on both the visibility and

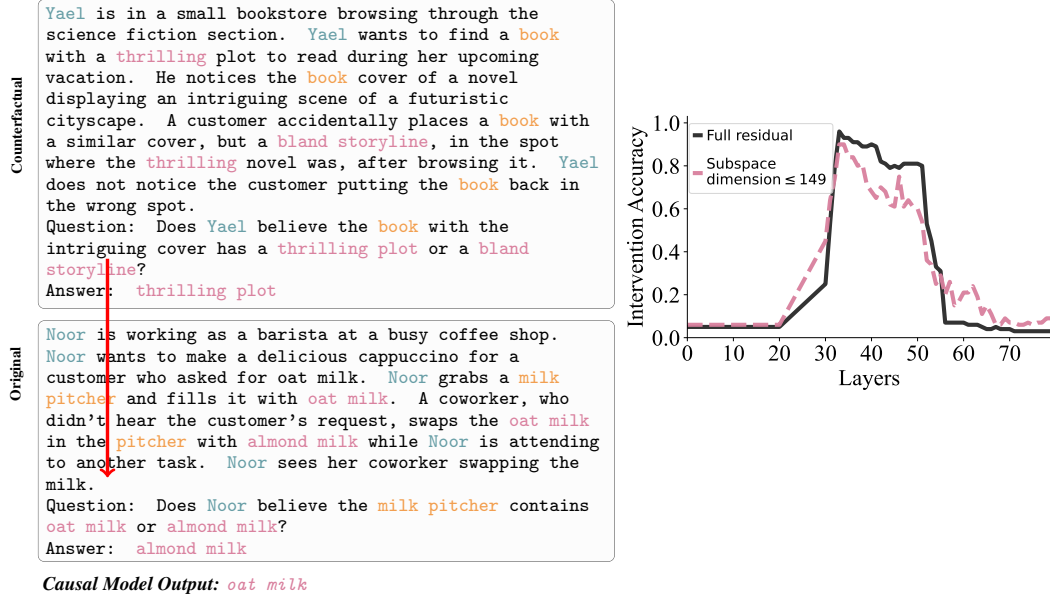


Figure 21: **Answer Lookback Pointer in BigToM**: This interchange intervention experiment modifies the pointer information (●) of the Answer lookback, thereby altering the subsequent QK-circuit to attend to the other state (e.g., *oat milk*) instead of the original one (e.g., *almond milk*). As a result, the model retrieves the token value corresponding to the other state to answer the question.

question sentences results in alignment across all layers. These results align with those found in the CausalToM setting shown in the Fig. 7.

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.

M 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.

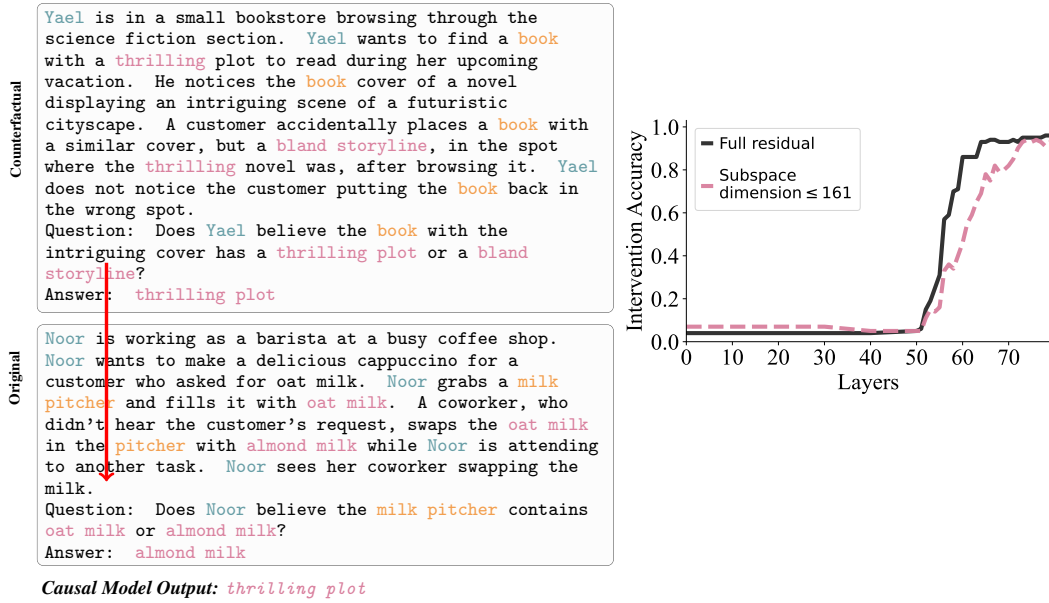


Figure 22: Answer Lookback Payload in BigToM: This interchange intervention experiment directly modifies the payload information (\blacktriangle) of the Answer lookback, which is fetched from the corresponding state tokens and predicted as the next token(s). Thus, replacing its value in the original 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.

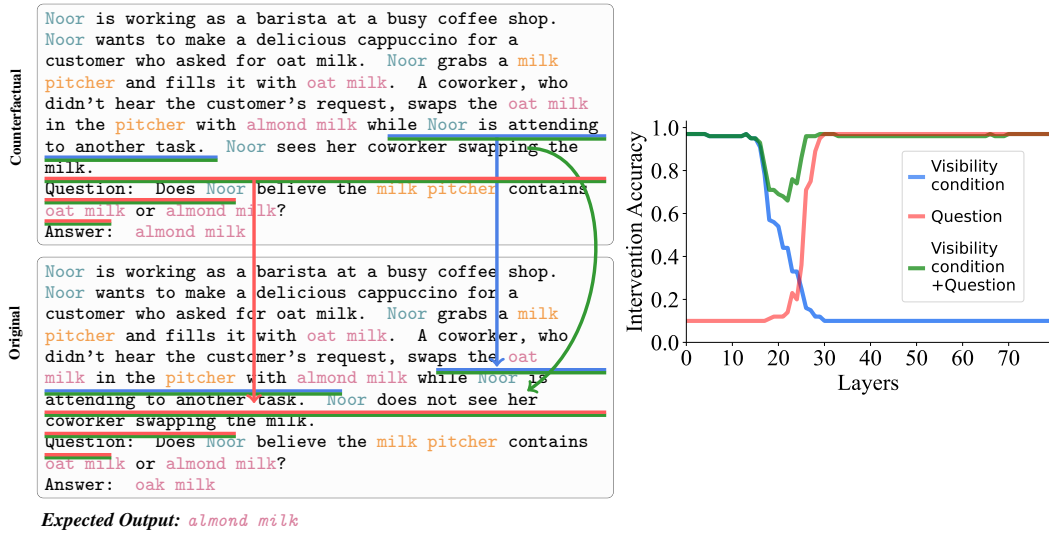


Figure 23: Visibility Lookback in BigToM: We perform three interchange interventions to establish the presence of the Visibility ID, which serves as both address and pointer information. When intervening at the source (\bullet)—i.e., the visibility sentence—both the address and pointer are updated, resulting 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 (\blacktriangle). However, intervening at both the visibility and question sentences results in alignment across all layers, as the address and pointer remain consistent throughout.

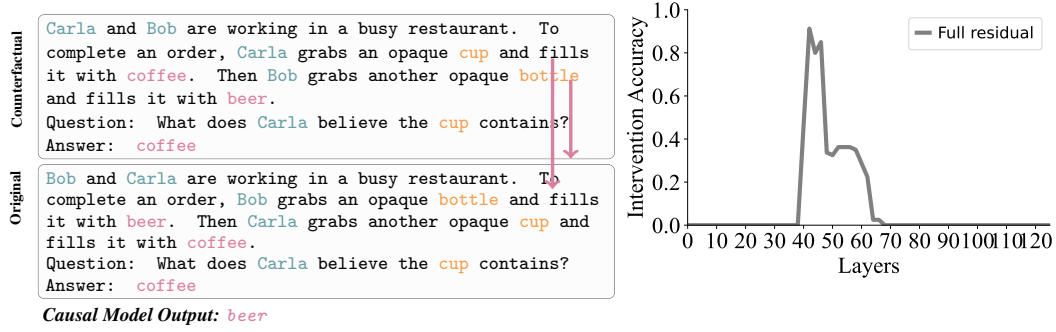


Figure 24: Payload and address of Binding lookback

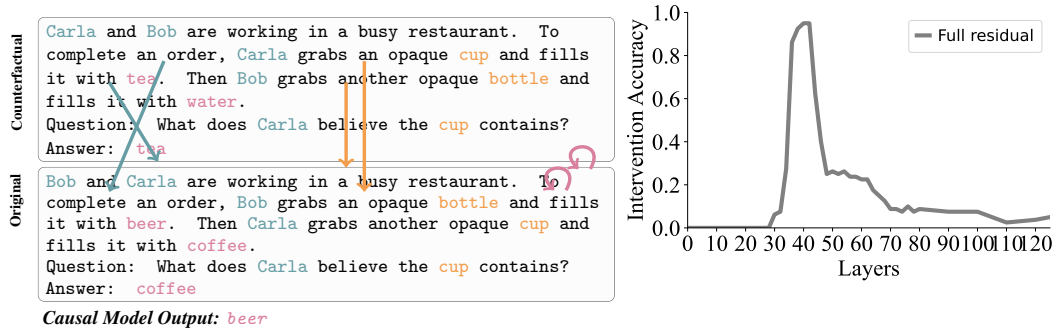


Figure 25: Source Information of Binding lookback

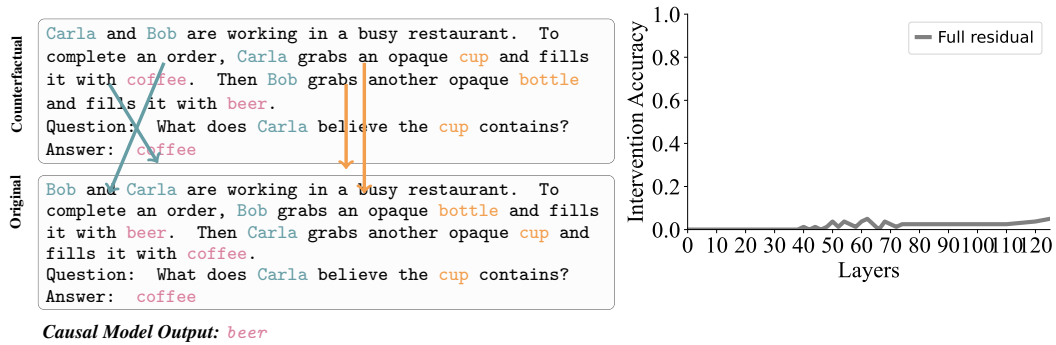


Figure 26: Source Information of Binding lookback without freezing address and payload

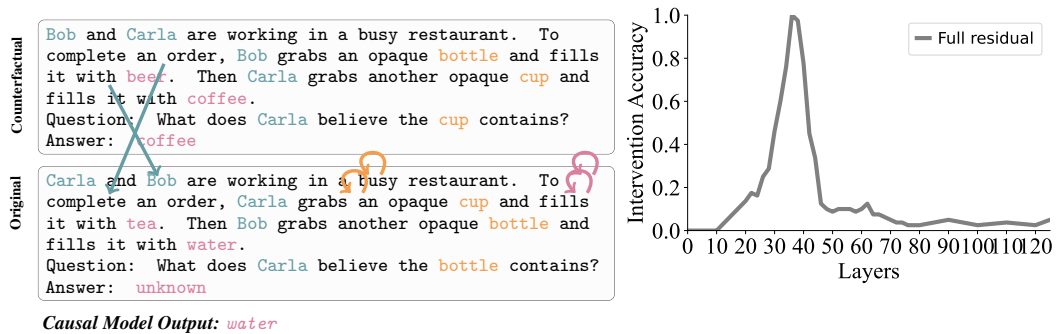


Figure 27: Character OI

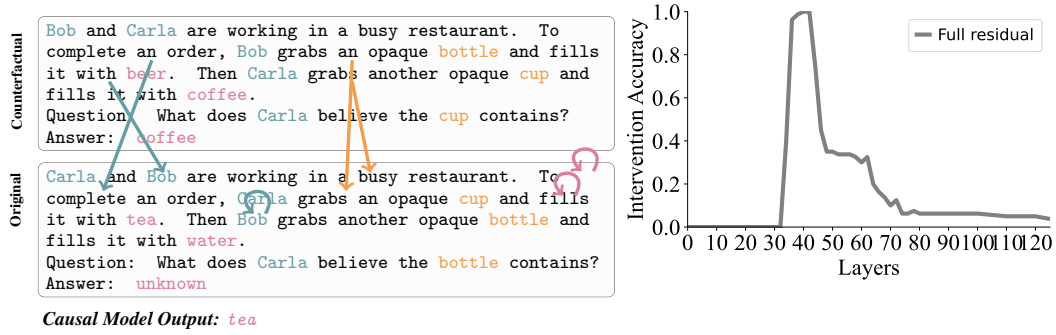


Figure 28: Object OI

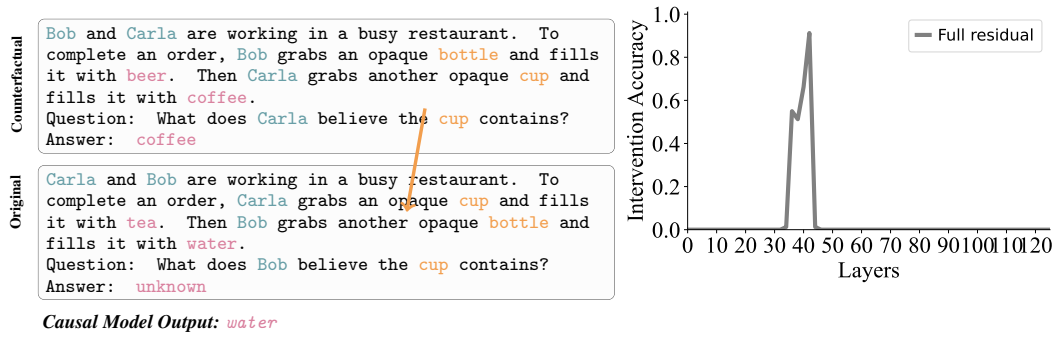


Figure 29: Query Object OI

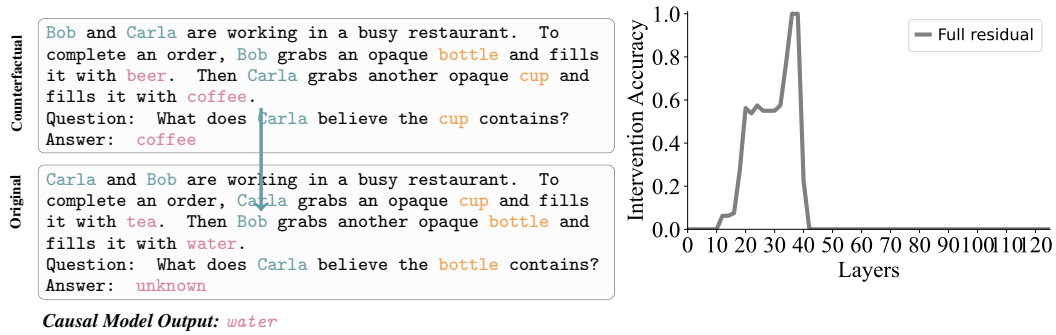


Figure 30: Query Character OI

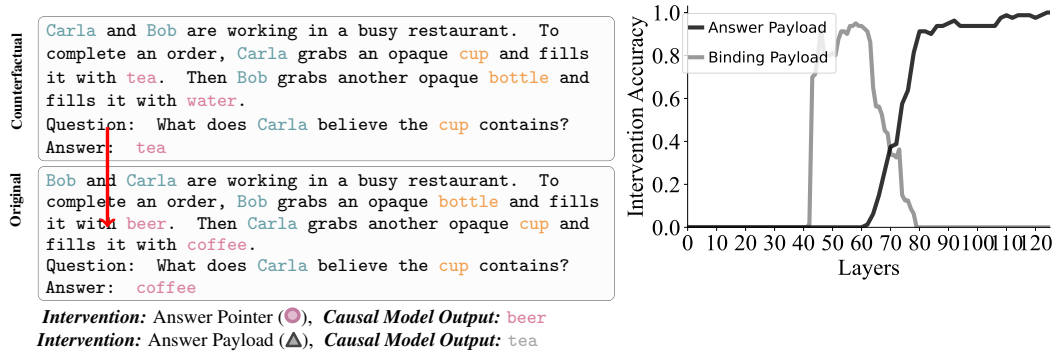


Figure 31: Answer Lookback Pointer and Payload

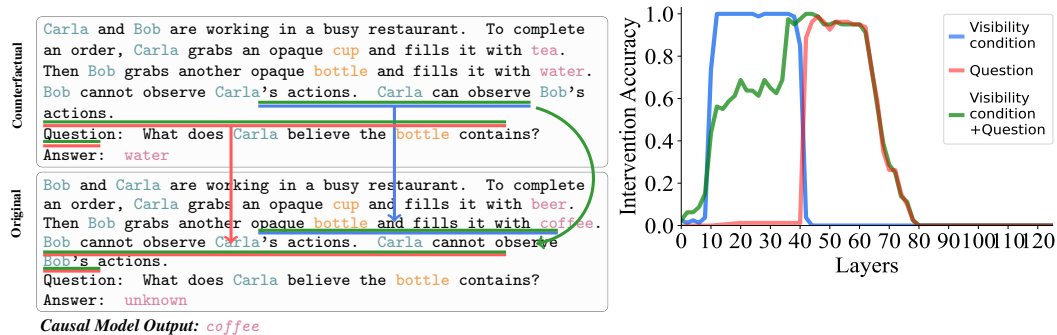


Figure 32: Visibility Lookback