LLMs use Lookbacks to solve simple Theory of Mind tasks

Nikhil Prakash, Natalie Shapira, Arnab Sen Sharma, Christoph Riedl,
Yonatan Belinkov, Tamar Rott Shaham, David Bau, Atticus Geiger

CausalToM Sample

No Visibility

Instruction: …
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:

Verifying ② and △ at final token

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

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

Intervention 1: Answer Pointer (●), Causal Model Output: beer
Intervention 2: Answer Payload (△), Causal Model Output: tea

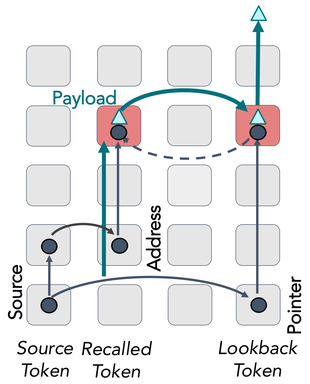
Verifying Visibility Lookback

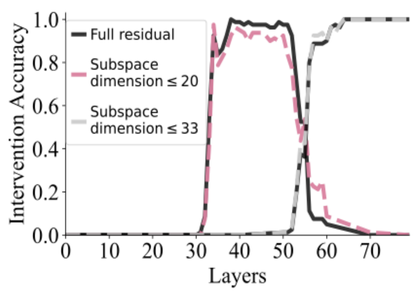
Carla and Bob are working in a busy restaurant. To complete an order, Carla grabs an opaque cup and fills it with tea. Then Bob grabs another opaque bottle and fills it with water. Bob cannot observe Carla's actions. Carla can observe Bob's actions.
Question: What does Carla believe the cup contains?
Answer: tea

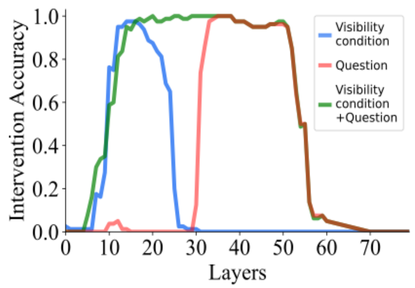
Karen and Max are working in a busy restaurant. To complete an order, Karen grabs an opaque flute and fills it with soda. Then Max grabs another opaque jar and fills it with coffee. Max cannot observe Karen's actions. Karen cannot observe Max's actions.
Question: What does Karen believe the jar contains?
Answer: **unknown**

Intervention 1: Source (●), Causal Model Output: coffee
Intervention 2: Payload (△), Causal Model Output: coffee
Intervention 2: Address & Pointer (●), Causal Model Output: coffee

The Lookback Mechanism



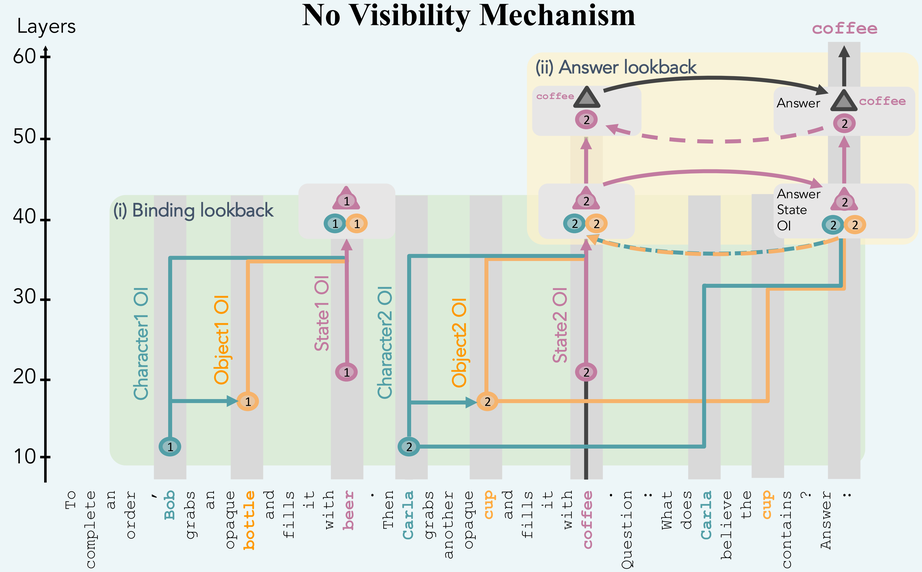




- LLMs form **ordering-based symbolic representations** to encode important contextual information.
- They perform binding operations, using **lookback mechanisms**, on the symbolic representations.
- Rather than relying on memorization, LLMs appear to use a **coherent algorithmic process** to solve theory-of-mind tasks.

belief.baulab.info

No Visibility Mechanism



With Visibility Mechanism