Understanding Social Reasoning in Language Models with Language Models

Kanishk Gandhi*, Jan-Philipp Franken*, Tobias Gerstenberg, Noah Goodman

Can language models attribute mental states to humans / agents?

	Sap+2022
Trott+ 2022	Ullman 03-16
Kosinski 02-04	Sap+ 04-03
Bubeck+ 03-23	Shapira+ 05-24
Moghaddam+ 04-26	Sclar+ 06-01
✅ ToM	❌ ToM

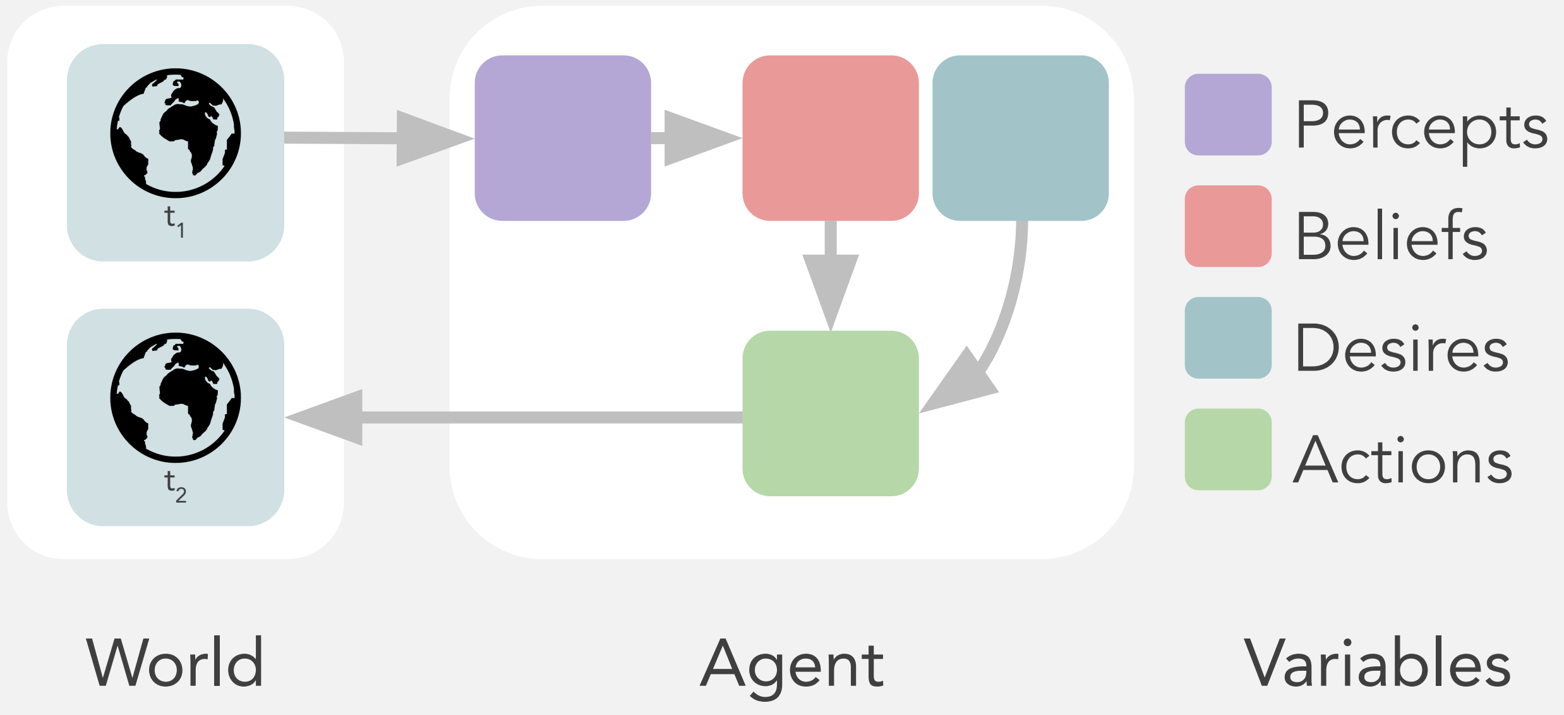
Problems with evaluations

Data quality   Dataset Leakage   Limited Test Items

Confounding Capabilities   No Rigorous Control Conditions

BigToM: Build an abstract causal template

Causal Template



x

Example Scenario

Noor is working as a barista at a busy coffee shop. Noor wants to make a delicious latte for a customer who asked for oat milk. Noor grabs a milk pitcher and fills it with oat milk. Noor believes that the milk pitcher contains oat milk.

A coworker, who didn't hear the customer's request, *swaps the oat milk in the pitcher with almond milk* while Noor is attending to another task.

Prior   Causal Event

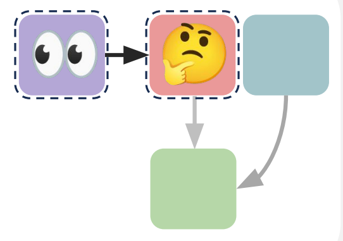
BigToM: Populate the template and generate conditions

Prompt Template

```
#          Generate          new          scenarios.
1. prior: {context} {desire} {action} {percept} {belief}
2. causal event: {causal event}
3. percepts: {percept true belief} {percept false belief}
4. actions: {action true belief} {action false belief}
5. questions: {belief question} {action question}
6. belief answers: {belief true belief} {belief false belief}
7. action answers: {action true belief} {action false belief}
```

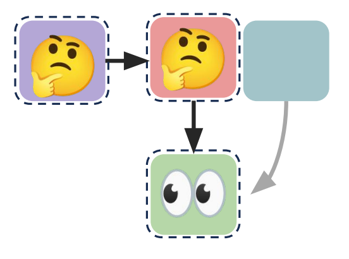
Conditions

```
# 1 forward belief true
prior: {context} {desire} {action} {percept}
{belief}
causal event: {causal event}
percept: {percept true belief}
question: {belief question}
```



Observed   Inferred

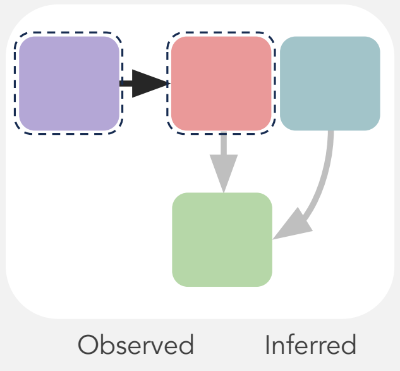
```
# 6 backward belief false
prior: {context} {desire} {action} {percept}
{belief}
causal event: {causal event}
percept: {action false belief}
question: {belief question}
```



25 conditions from each populated template

Human Performance

Forward Belief



Causal Inference

Percepts:

True Belief
Noor sees her coworker swapping the milk.

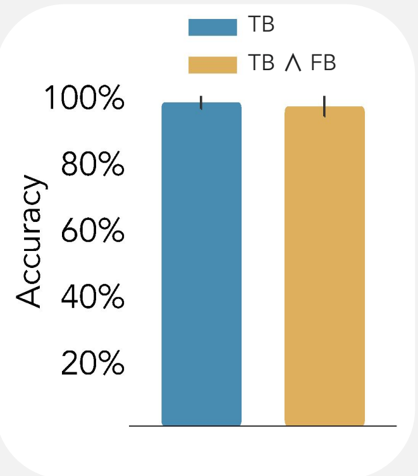
False Belief
Noor does not see her coworker swapping the milk.

Question: What does Noor believe?

Answer Options:

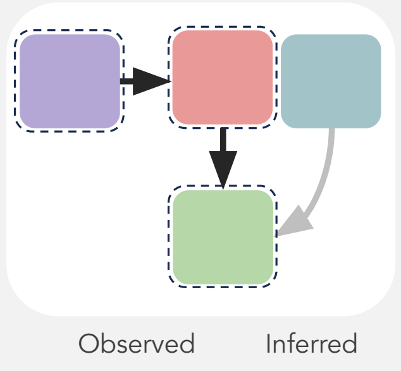
True Belief
Noor believes that the milk pitcher contains almond milk.

False Belief
Noor believes that the milk pitcher contains oat milk.



Causal Inference   Percepts   Question   Answer Options   Human Performance

Forward Action



Percepts:

True Belief
Noor sees her coworker swapping the milk.

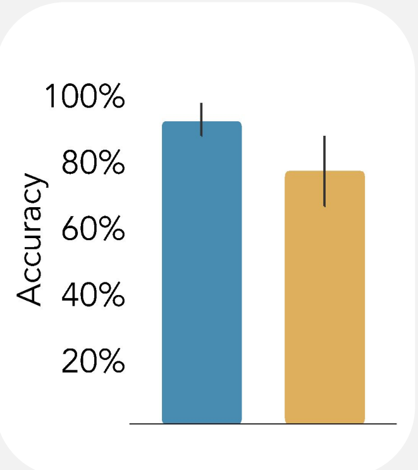
False Belief
Noor does not see her coworker swapping the milk.

Question: What will Noor do?

Answer Options:

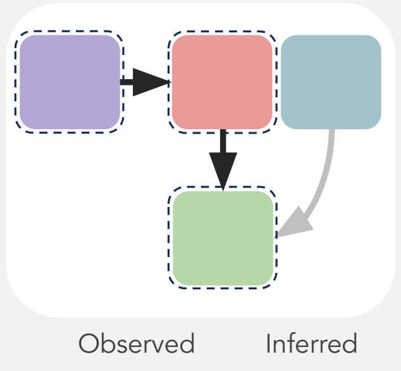
True Belief
Noor opens the fridge again and reaches for oat milk.

False Belief
Noor makes the latte using the milk in the pitcher.



Causal Inference   Percepts   Question   Answer Options   Human Performance

Backward Belief



Percepts:

True Belief
Noor opens the fridge again and reaches for oat milk.

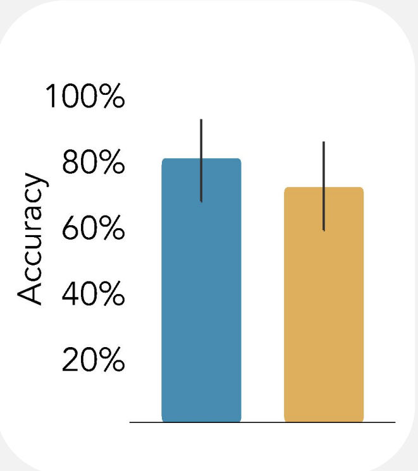
False Belief
Noor makes the latte using the milk in the pitcher.

Question: What does Noor believe?

Answer Options:

True Belief
Noor believes that the milk pitcher contains almond milk.

False Belief
Noor believes that the milk pitcher contains oat milk.



Causal Inference   Percepts   Question   Answer Options   Human Performance

Quality of generated data

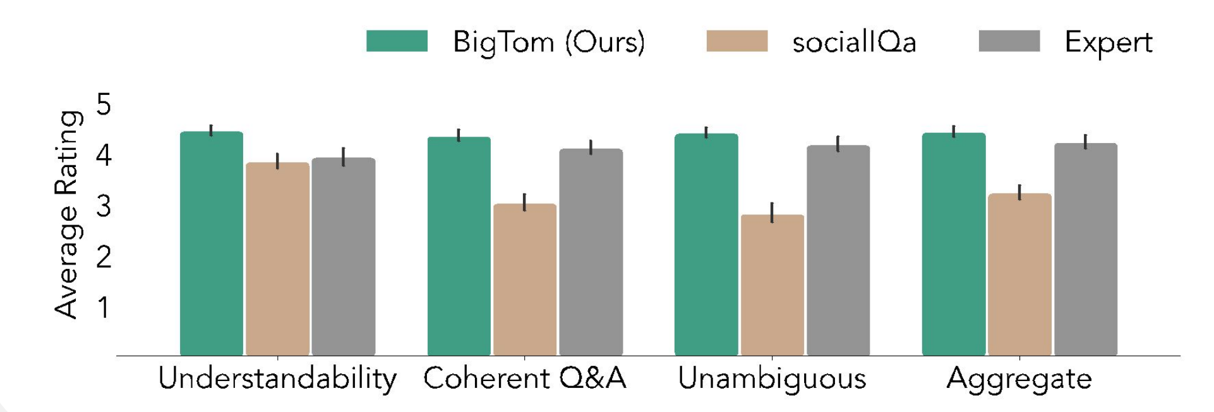
Questions

Understandability:
The story is easy to understand.
(1 strongly disagree–5 strongly agree)

Coherent question and answer:
The question and answers are relevant and clear in relation to the story.
(1 strongly disagree–5 strongly agree)

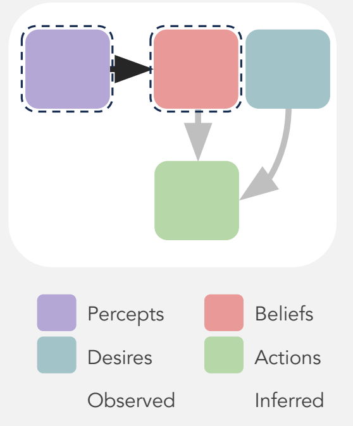
Unambiguous:
The correct answer to the question is clear and unambiguous.
(1 strongly disagree–5 strongly agree)

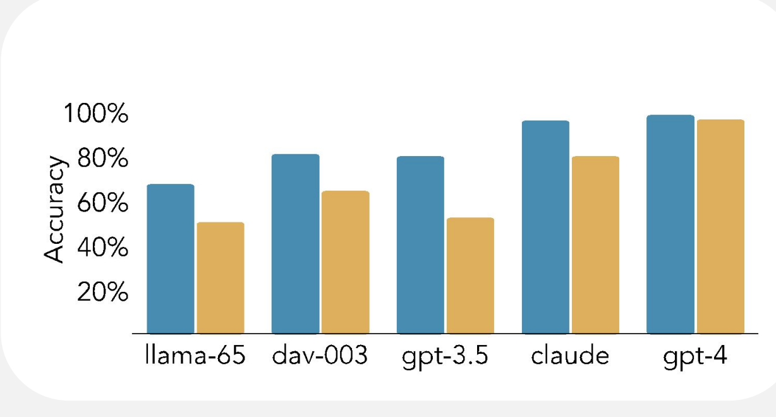
Human Ratings

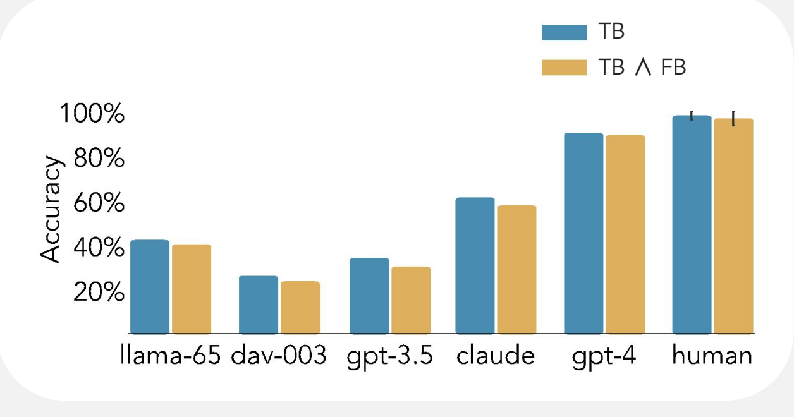


Model Performance

[a] Forward Belief

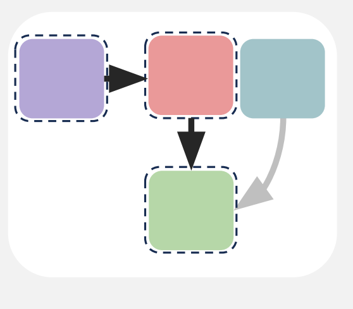


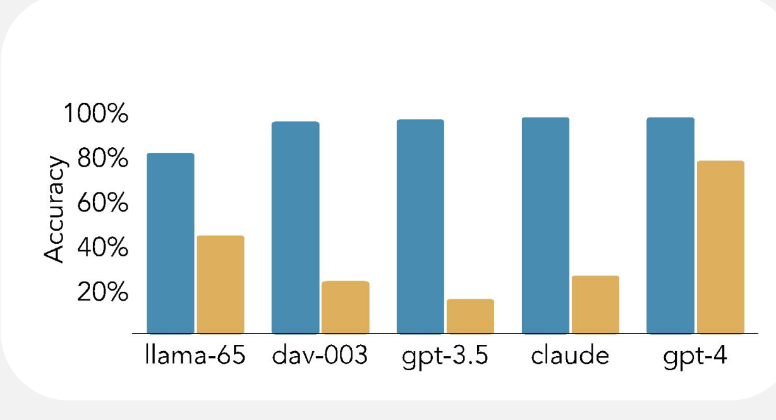


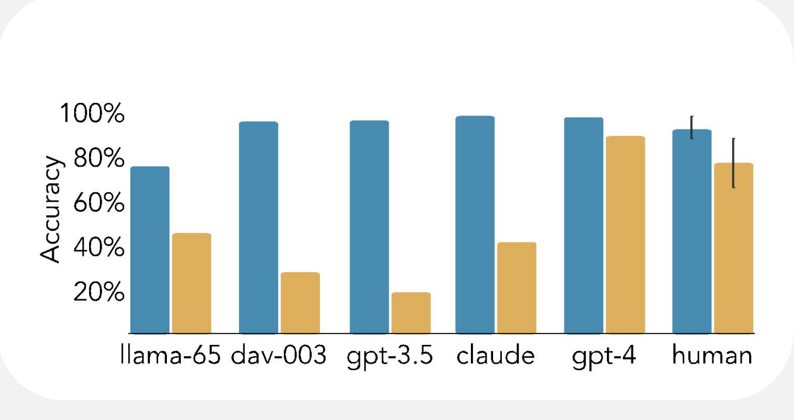


Causal Inference   0-shot Performance (without initial belief)   0-shot Performance (with initial belief)

[b] Forward Action

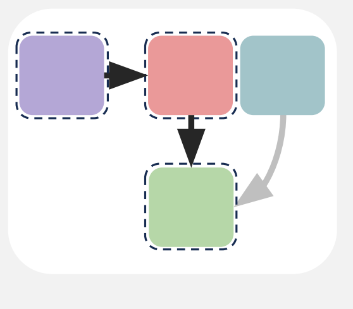


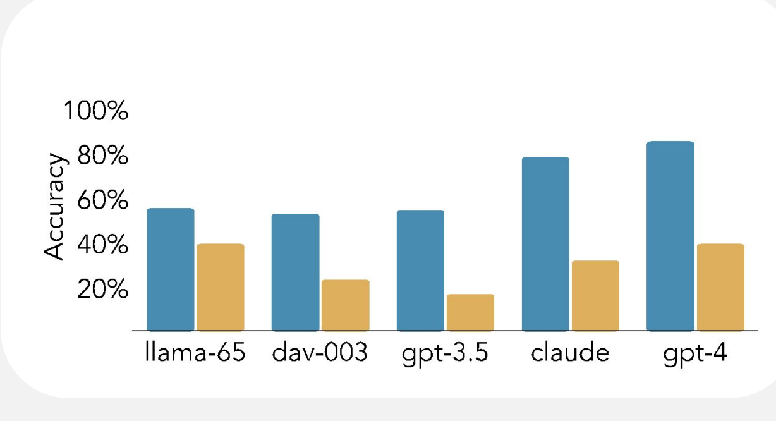


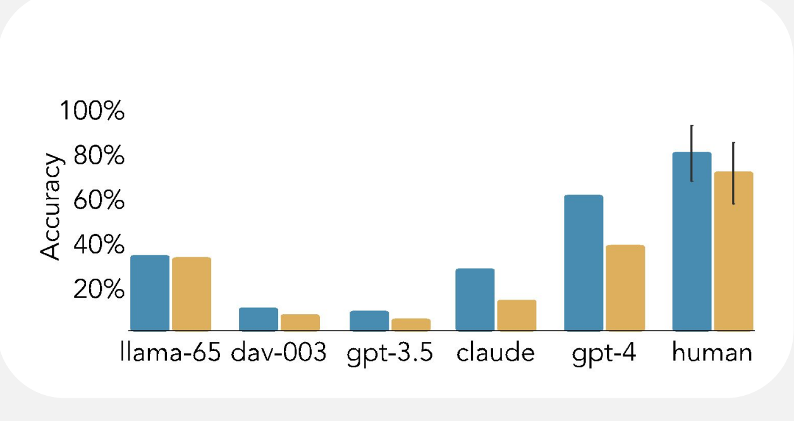


Causal Inference   0-shot Performance (without initial belief)   0-shot Performance (with initial belief)

[c] Backward Belief







Causal Inference   0-shot Performance (without initial belief)   0-shot Performance (with initial belief)

Advantages of BigToM

Scaleable   Smaller Risk of Leakage   Cheap to Generate

Measures of Task Validity   Careful Control Conditions