



CaT-Bench: Benchmarking Language Model Understanding of Causal and Temporal Dependencies in Plans



Yash Kumar Lal*



Vanya Cohen*



Nathanael Chambers



Niranjan
Balasubramanian



Raymond Mooney



Plans are ubiquitous

Goal: Bake Almond and Chocolate Cake

...

Step 6: Stir in ground almonds.
Step 7: Add half flour and half milk.
Step 8: Use wooden spoon to stir.

...

Step 12: Whip cream till stiff peaks

...

Goal: Navigate to Empire State Building

...

Step 3: Enter the 72 St station.
Step 4: Take Q train downtown.
Step 5: Get off at 34 St station.

...

Step 8: Walk east to the building.

...

Goal: Study Stock Market to Sell Gold Options

...

Step 2: Collate gold ETF history.
Step 3: Study for patterns in it.
Step 4: Find the highs and lows.

...

Step 7: Extrapolate trends...

...

LLMs generate fluent plans



plan a day and a half trip to yellowstone national park from salt lake city in early may



Day 1: Leave Salt Lake City early in the morning to make the most of your time at Yellowstone National Park. The drive is approximately 320 miles (515 km) and takes about 5-6 hours, depending on traffic and road conditions.

Here's a suggested itinerary for your day and a half trip:

Morning:

- Depart from Salt Lake City around 6:00 AM.
- Drive through scenic Idaho, passing by Craters of the Moon National Monument and Preserve if time allows.

How do we check correctness of these plans?

Plans contain dependencies

Goal: Bake Almond and
Chocolate Cake

...

Step 6: Stir in ground almonds.

Step 7: Add half flour and half milk.

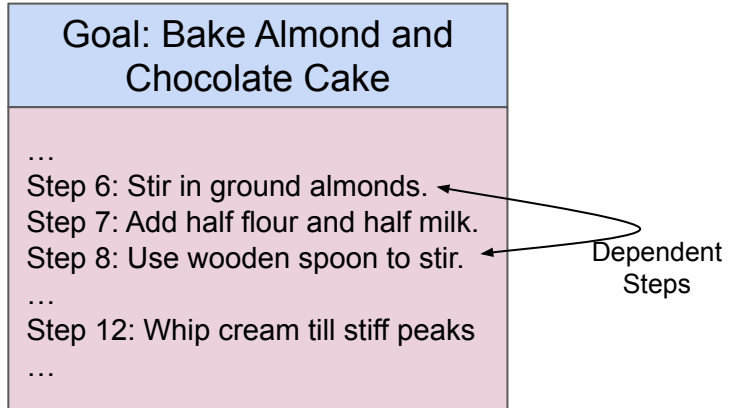
Step 8: Use wooden spoon to stir.

...

Step 12: Whip cream till stiff peaks

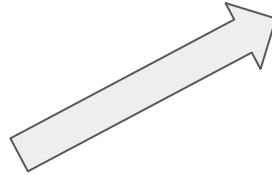
...

Plans contain dependencies



Plans contain dependencies

Goal: Bake Almond and Chocolate Cake
...
Step 6: Stir in ground almonds.
Step 7: Add half flour and half milk.
Step 8: Use wooden spoon to stir.
...
Step 12: Whip cream till stiff peaks
...

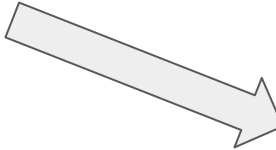
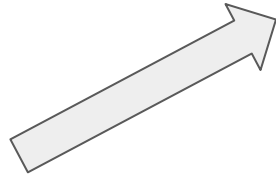


Q: Must Step 6 happen before Step 8?
Preconditions
A: Yes, all ingredients have to be in bowl before stirring

Questions about dependent steps

Plans contain dependencies

Goal: Bake Almond and Chocolate Cake
...
Step 6: Stir in ground almonds.
Step 7: Add half flour and half milk.
Step 8: Use wooden spoon to stir.
...
Step 12: Whip cream till stiff peaks
...



Q: Must Step 6 happen before Step 8?
Preconditions
A: Yes, all ingredients have to be in bowl before stirring

Questions about dependent steps

Q: Must Step 7 happen after Step 6?
Parallel Steps
A: No, almonds can be added after flour and milk

Questions about non-dependent steps

Prior Work

Goal: Bake Almond and
Chocolate Cake

...

Step 6: Stir in ground almonds.

Step 7: Add half flour and half milk.

Step 8: Use wooden spoon to stir.

...

Step 12: Whip cream till stiff peaks

...

Prior Work

Goal: Bake Almond and Chocolate Cake

...

Step 6: Stir in ground almonds.

Step 7: Add half flour and half milk.

Step 8: Use wooden spoon to stir.

...

Step 12: Whip cream till stiff peaks

...

Before Step 6, what is the state of almonds?

ground

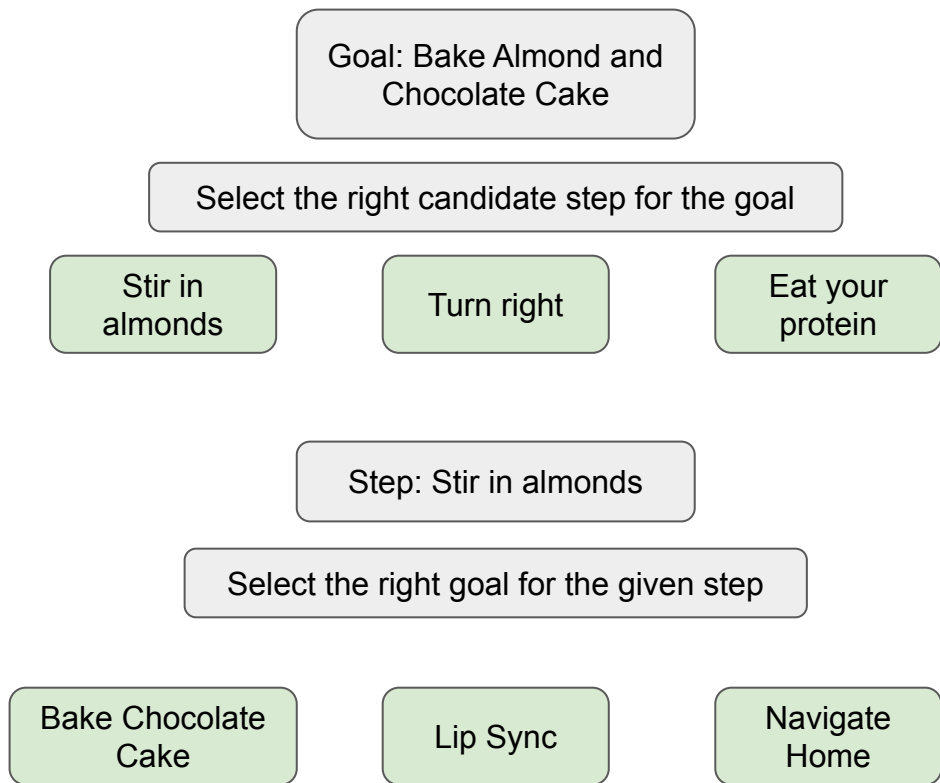
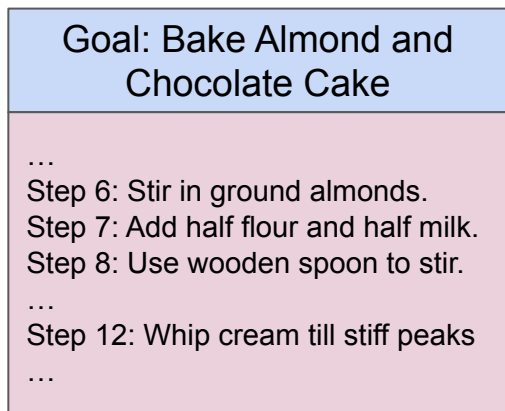
After Step 7, what is the state of the flour?

mixed

After Step 12, what is the state of the cream?

whipped

Prior Work



Prior Work

Goal: Bake Almond and
Chocolate Cake

...
Step 6: Stir in ground almonds.
Step 7: Add half flour and half milk.
Step 8: Use wooden spoon to stir.
...
Step 12: Whip cream till stiff peaks
...

Goal: Bake Almond and
Chocolate Cake

If you have non-stick utensils, what do you do after Step 7?

Use wooden
spoon to stir

Use your metal
spatula to stir

Creating 🐱 CaT-Bench

Goal: Bake Almond and
Chocolate Cake

...

Step 6: Stir in ground almonds.

Step 7: Add half flour and half milk.

Step 8: Use wooden spoon to stir.

...

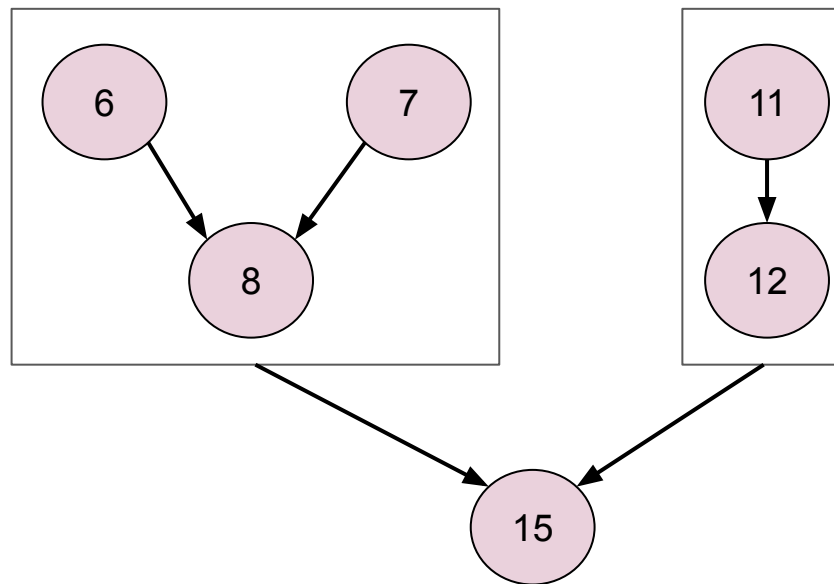
Step 12: Whip cream till stiff peaks

...

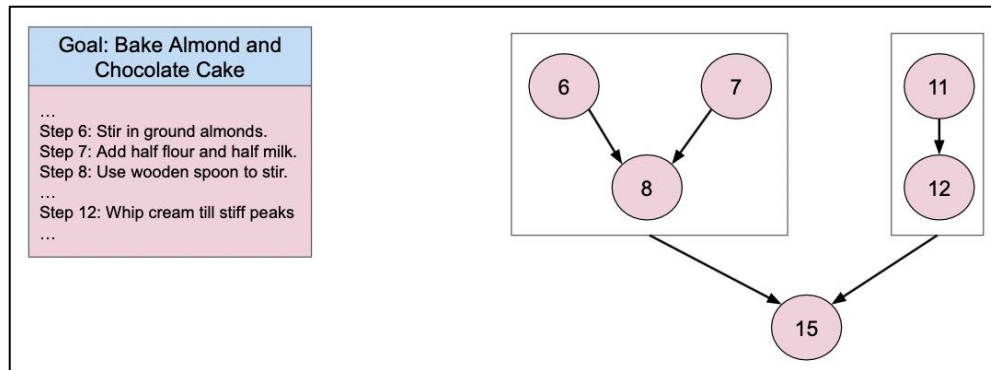
Creating 🐱 CaT-Bench

Goal: Bake Almond and
Chocolate Cake

...
Step 6: Stir in ground almonds.
Step 7: Add half flour and half milk.
Step 8: Use wooden spoon to stir.
...
Step 12: Whip cream till stiff peaks
...



Creating 🐱 CaT-Bench



710 Q: Must Step 6 happen before Step 8?
710 Q: Must Step 8 happen after Step 6?

1420

Questions about dependent steps

710 Q: Must Step 12 happen after Step 6?
710 Q: Must Step 6 happen before Step 12?

1420

Questions about non-dependent steps

Benchmarking Models

- GPT-3.5
- GPT-4-Turbo
- GPT-4o
- Llama-3-8B
- Claude-3.5-Sonnet
- Gemini-1.0-Pro
- Gemini-1.5-Pro
- Gemini-1.5-Flash

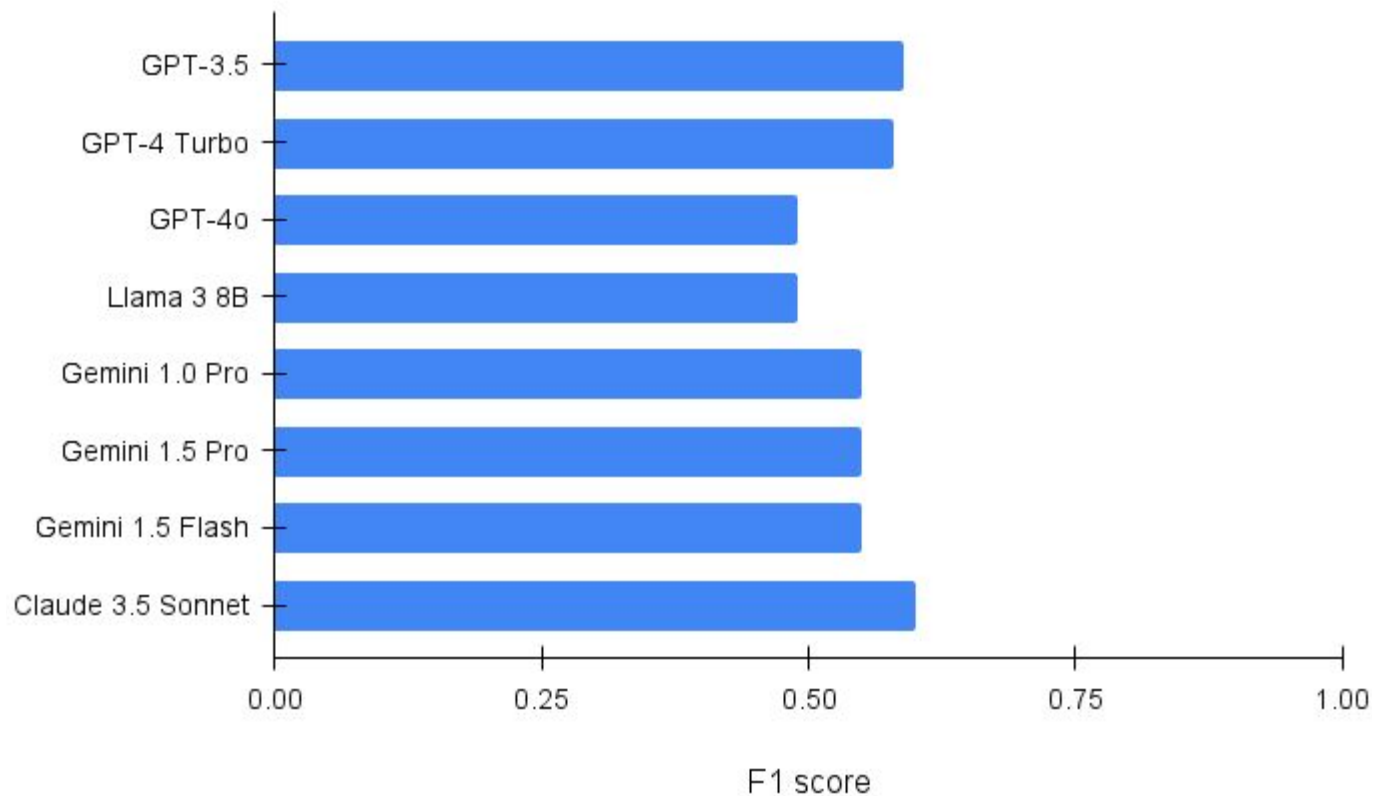


Experiments - Answer Only (A)

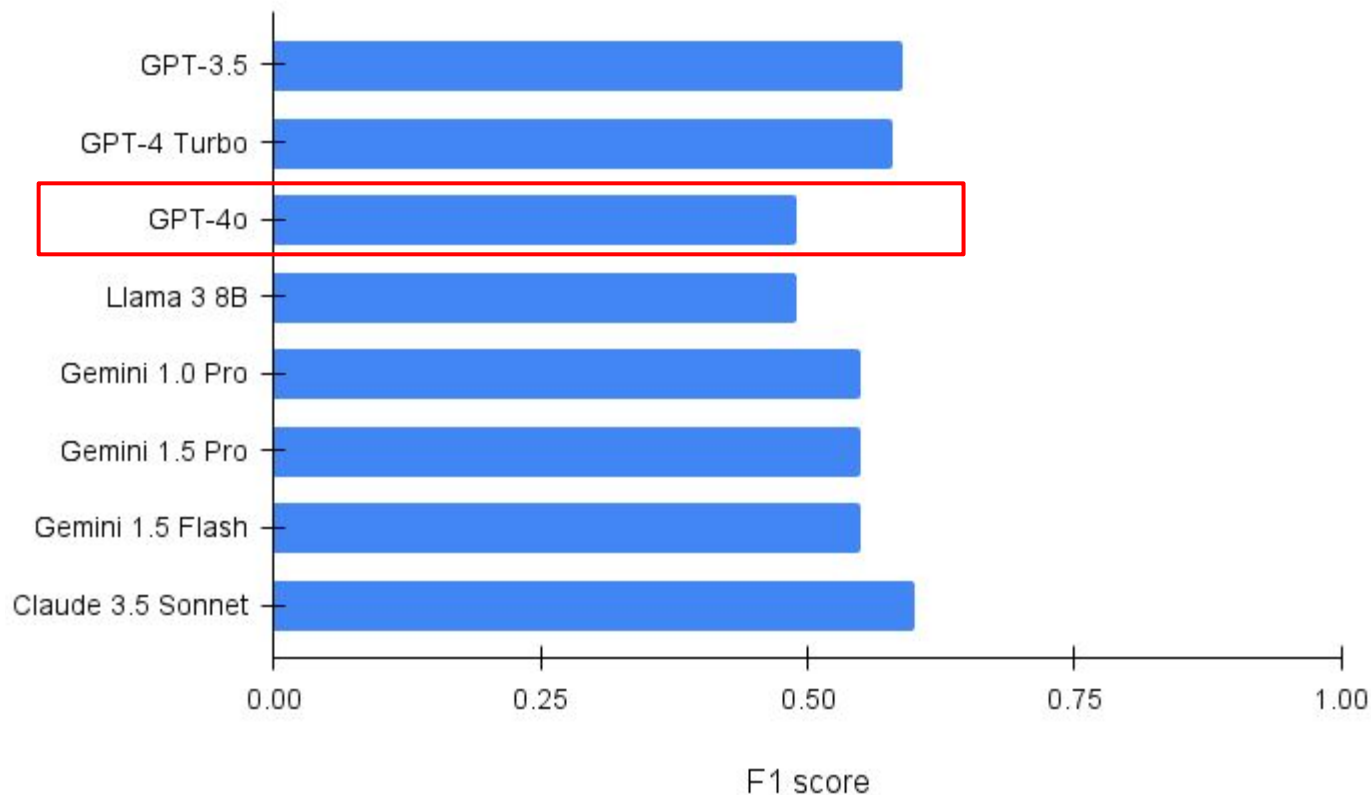
- Binary Prediction

Answer-only (A)	<p>Given a goal, a procedure to achieve that goal and a question about the steps in the procedure, you are required to answer the question in one sentence.</p> <p>Goal: {title}</p> <p>Procedure: {procedure}</p> <p>Must Step {i} happen before Step {j}? Select between yes or no</p>
-----------------	--

Models Struggle at Predicting Step Order



Models Struggle at Predicting Step Order

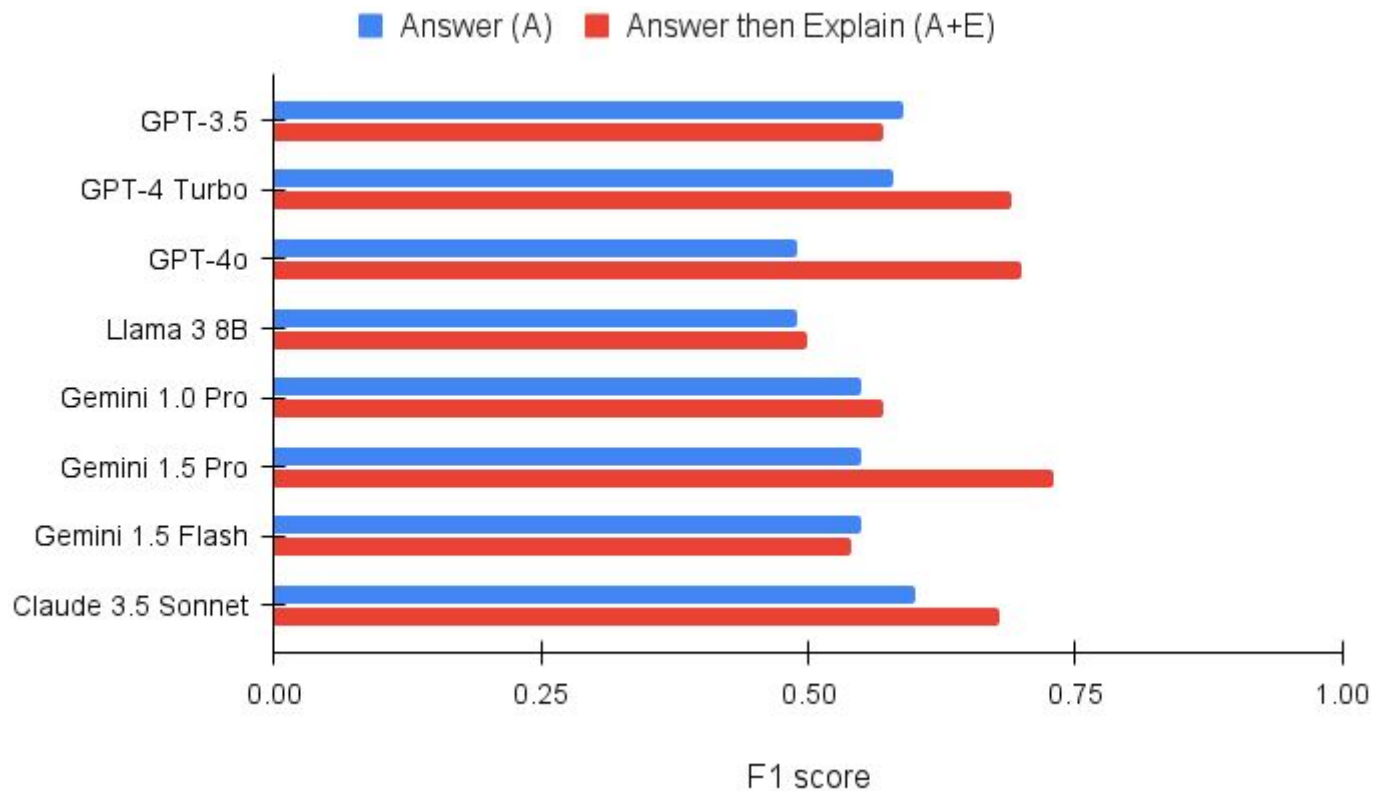


Experiments - Answer + Explanation (A+E)

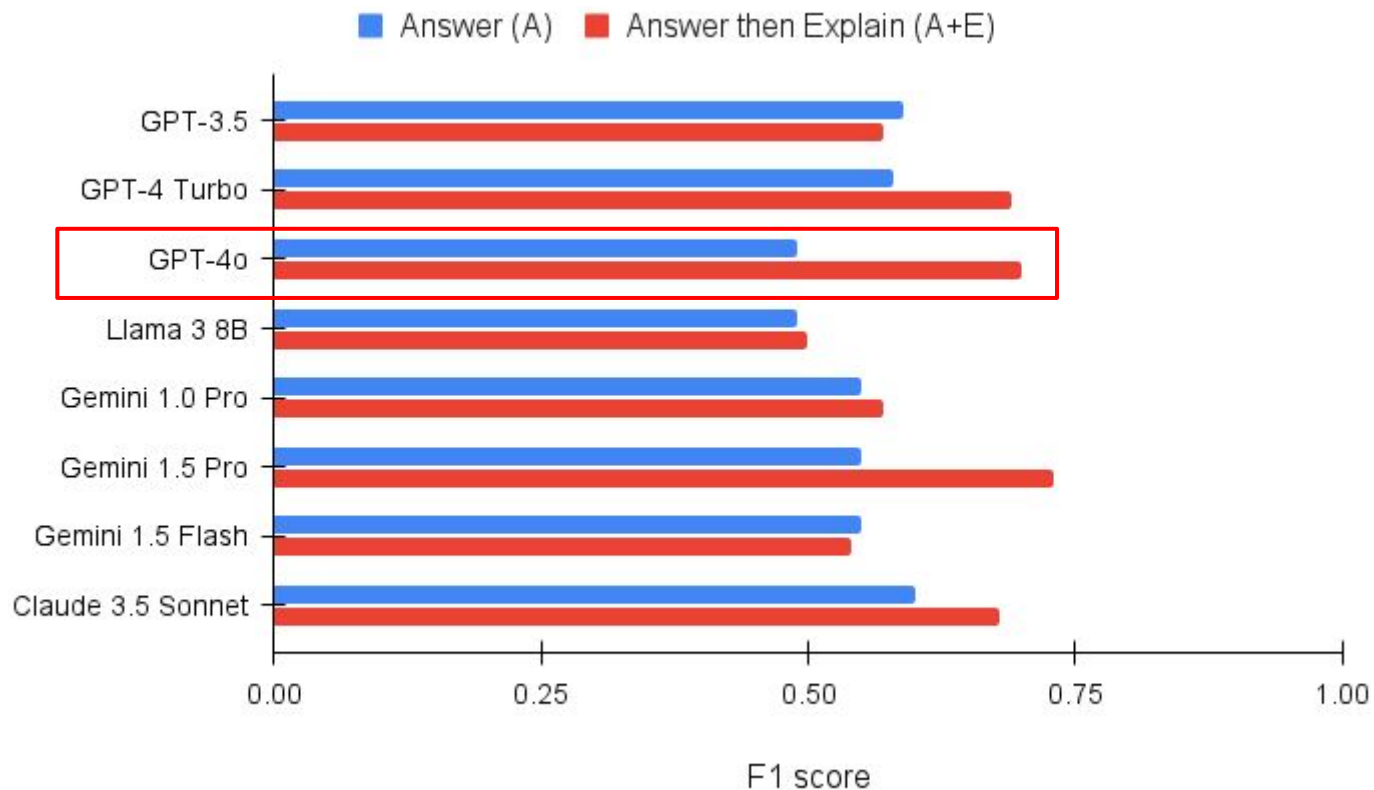
- Binary Prediction and Post-hoc Explanation

<p>Answer + Explanation (A+E)</p>	<p>Given a goal, a procedure to achieve that goal and a question about the steps in the procedure, you are required to answer the question in one sentence.</p> <p>Goal: {title}</p> <p>Procedure: {procedure}</p> <ol style="list-style-type: none">1. Must Step {i} happen before Step {j}? Select between yes or no2. Explain why or why not. <p>Format your answer as JSON with the key value pairs "binary_answer": "yes/no answer to Q1", "why_answer": "answer to Q2"</p>
---------------------------------------	---

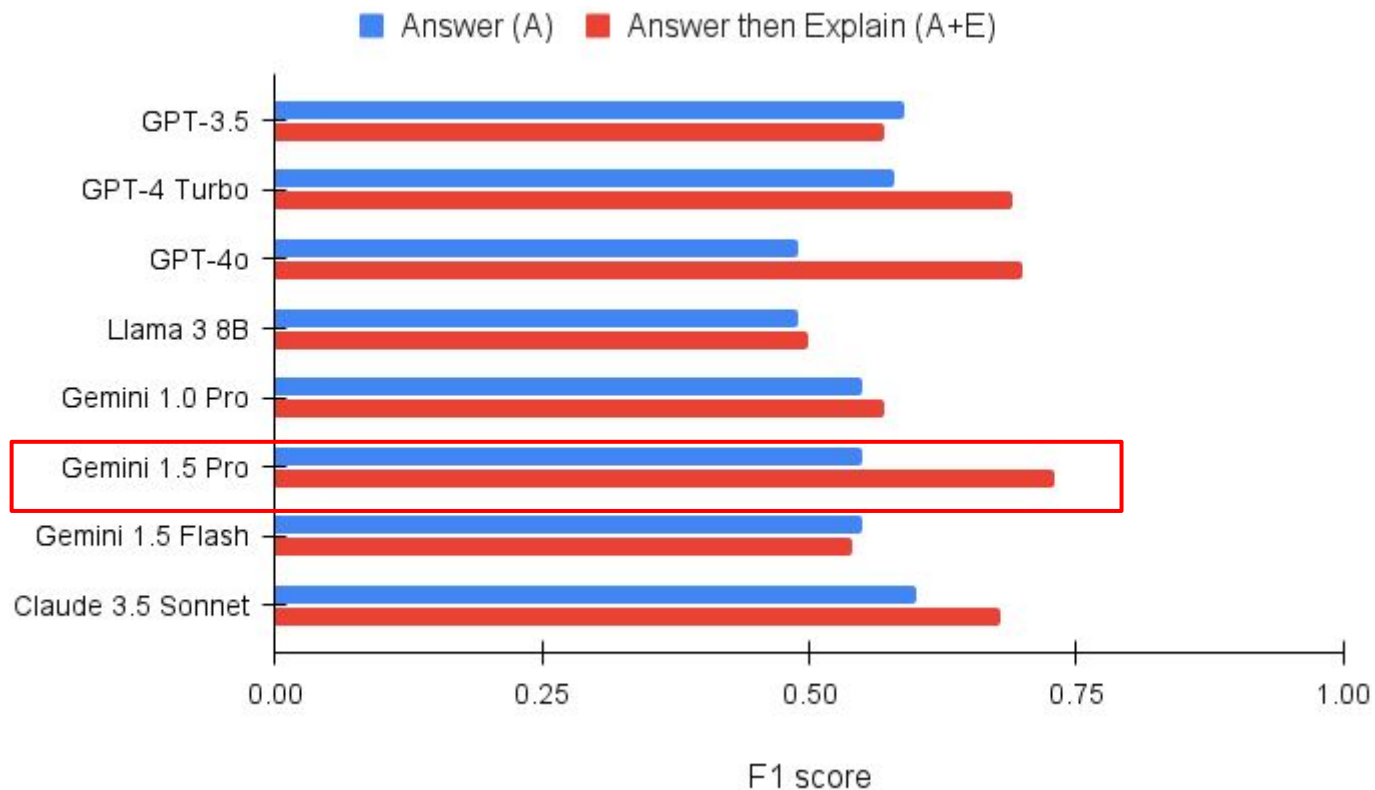
Generating explanations help!



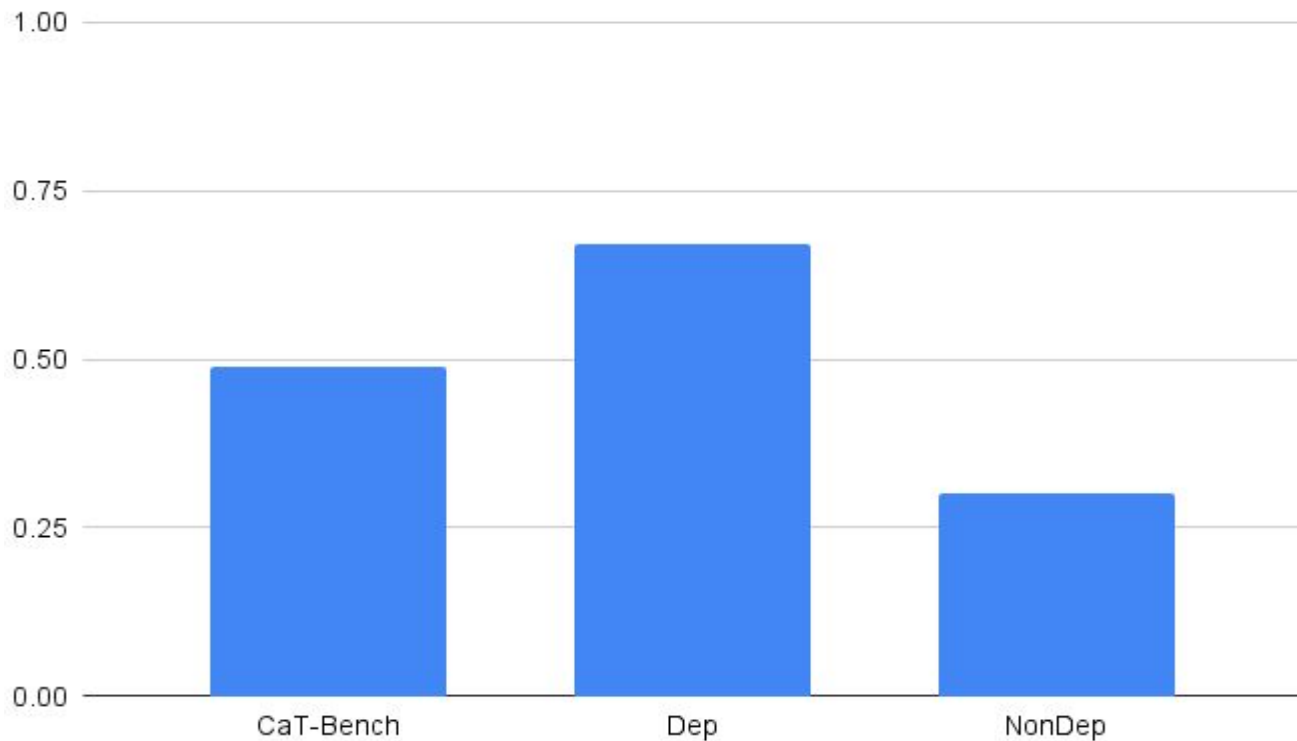
Generating explanations help!



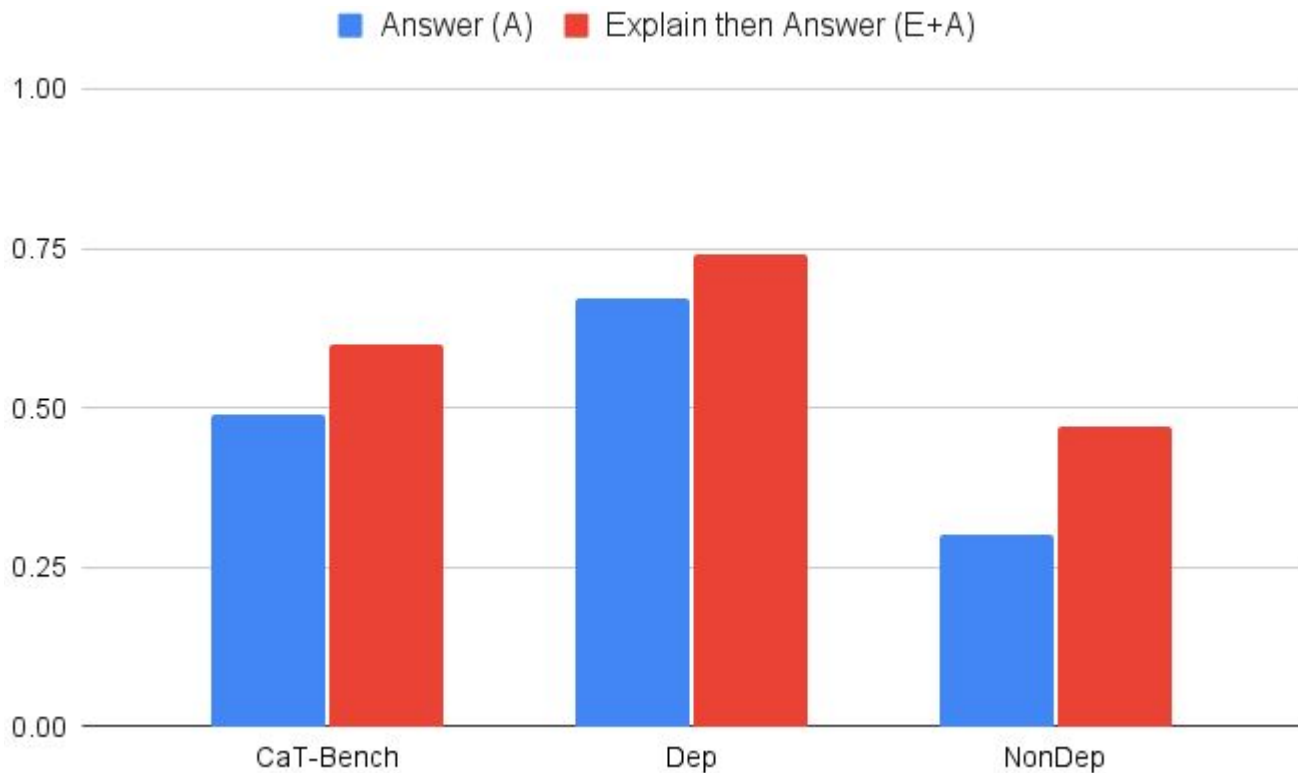
Generating explanations help!



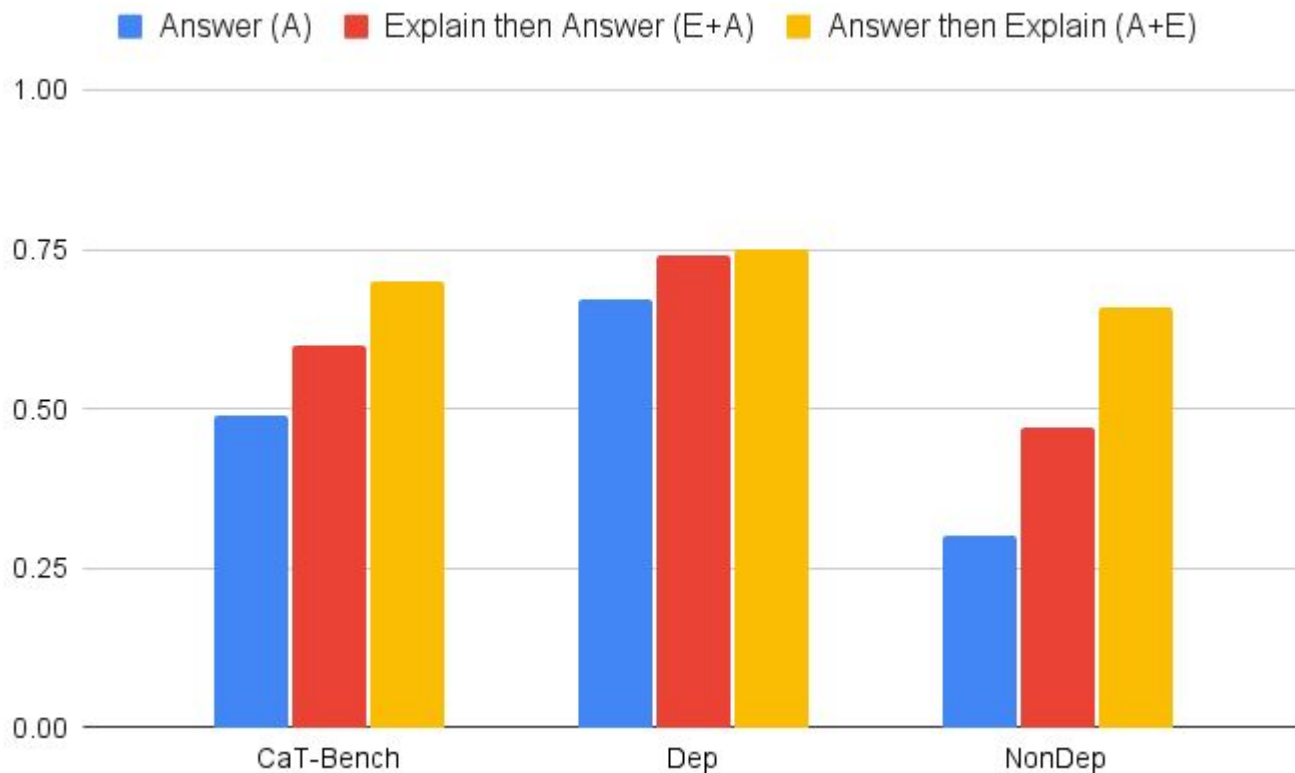
Intermediate Reasoning or Post-Hoc Explanation?



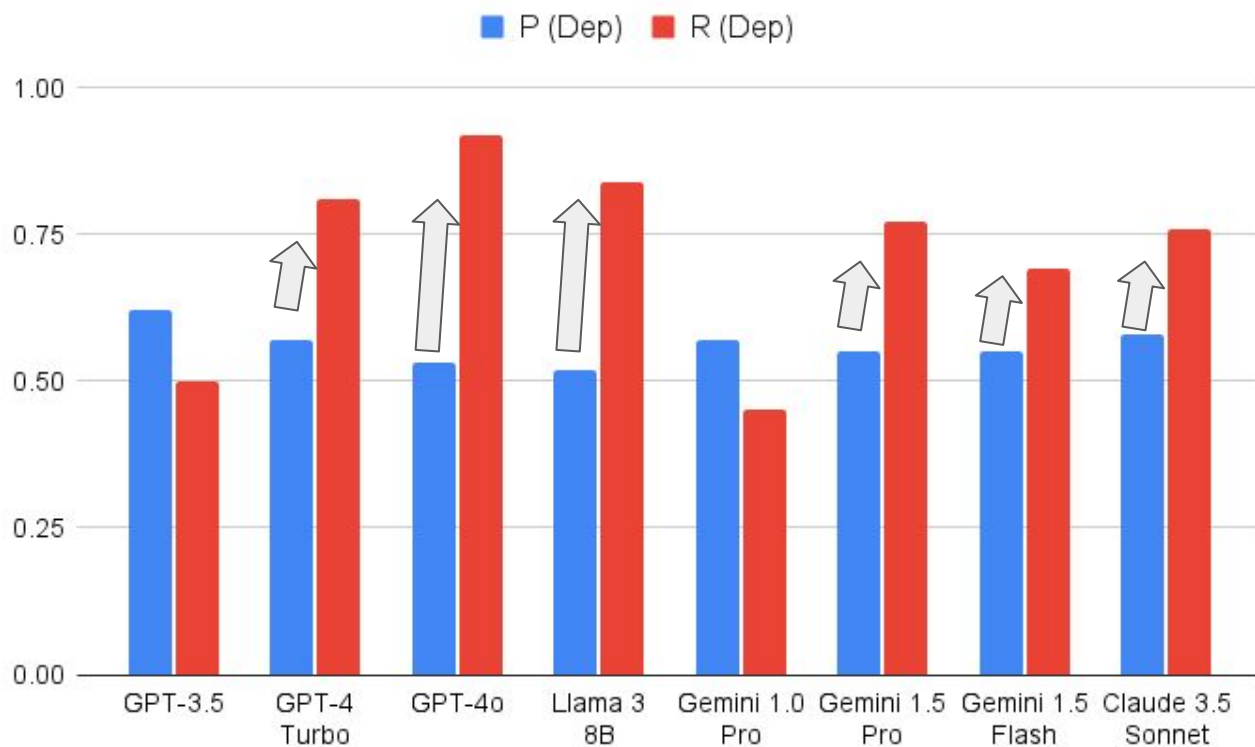
Chain of Thought Improves over Zero-Shot



Post-Hoc Explanations are Better than Chain of Thought



Models are Biased towards Predicting Dependence



How Good are the Generated Explanations?

Model Explanation: All ingredients must be in the bowl to be mixed well together.

The answer contains the relevant details to address the question.

Likert Scale

Strongly Disagree
(1)

Disagree
(2)

Neutral
(3)

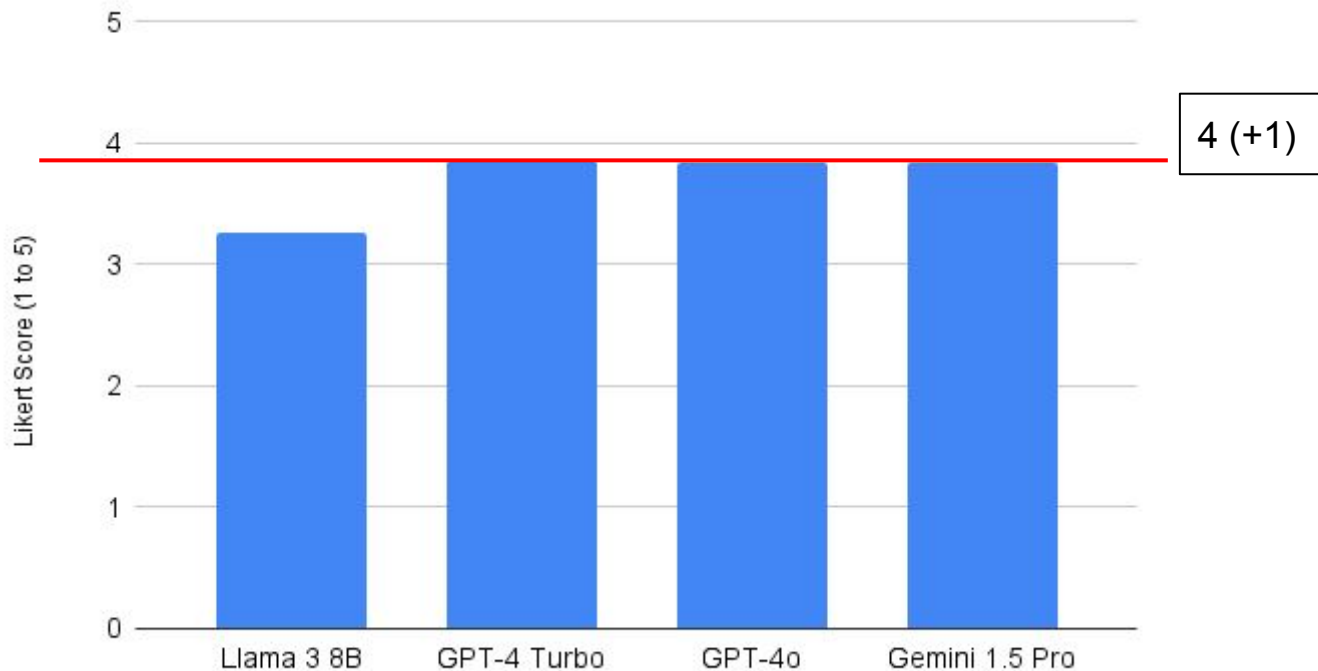
Agree
(4)

Strongly Agree
(5)

Average Likert Score



How Good are the Generated Explanations?



How Good are the Generated Faithful Explanations?

Model Answer: No

Model Explanation: All ingredients must be in the bowl to be mixed well together.

The answer contains the relevant details to address the question.

Likert Scale

Strongly Disagree
(1)

Disagree
(2)

Neutral
(3)

Agree
(4)

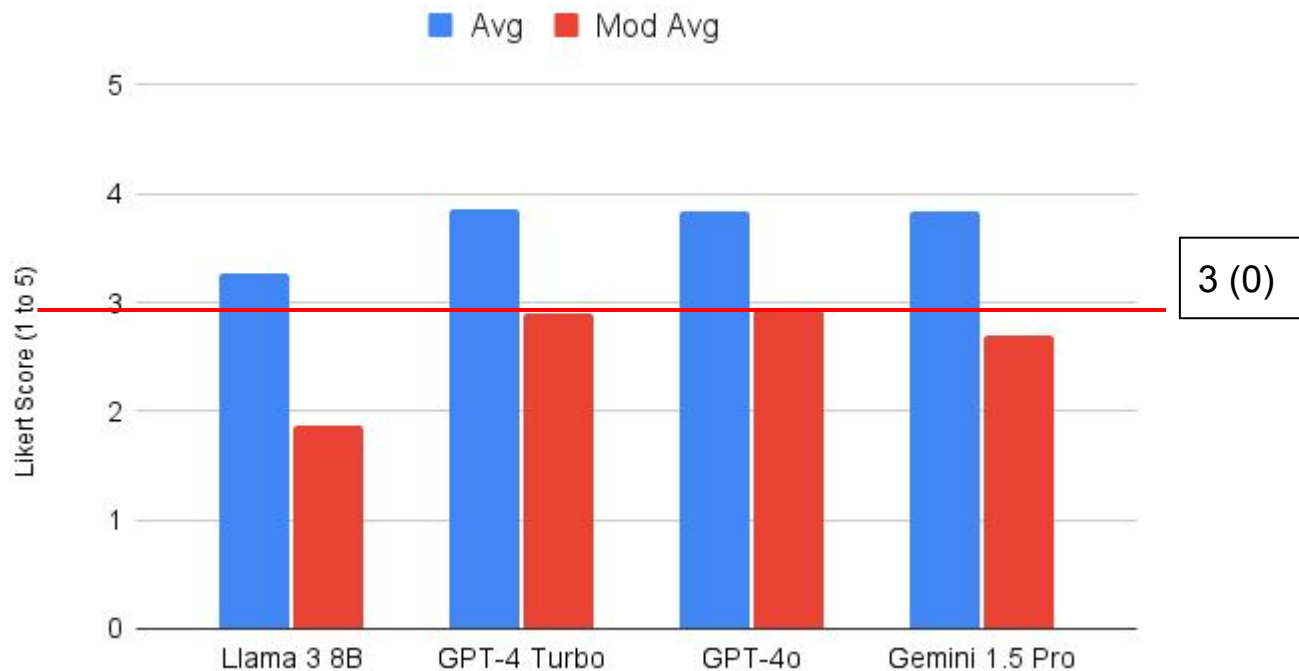
Strongly Agree
(5)

Average Likert Score



= 0

How Good are the Generated Faithful Explanations?



How Robust are Models: Temporal Consistency

Goal: Bake Almond and
Chocolate Cake

...
Step 6: Stir in ground almonds.
Step 7: Add half flour and half milk.
Step 8: Use wooden spoon to stir.
...
Step 12: Whip cream till stiff peaks
...

Q: Must Step 6 happen
before Step 8?

Before

Q: Must Step 8 happen
after Step 6?

After

The model answer to the 'before' and 'after' questions should be the same.

How Robust are Models: Order Contrastive Consistency

Goal: Bake Almond and
Chocolate Cake

...

Step 6: Stir in ground almonds.

Step 7: Add half flour and half milk.

Step 8: Use wooden spoon to stir.

...

Step 12: Whip cream till stiff peaks

...

How Robust are Models: Order Contrastive Consistency

Goal: Bake Almond and Chocolate Cake

...

Step 6: Stir in ground almonds.

Step 7: Add half flour and half milk.

Step 8: Use wooden spoon to stir.

...

Step 12: Whip cream till stiff peaks

...



Goal: Bake Almond and Chocolate Cake

...

Step 6: Add half flour and half milk.

Step 7: Stir in ground almonds.

Step 8: Use wooden spoon to stir.

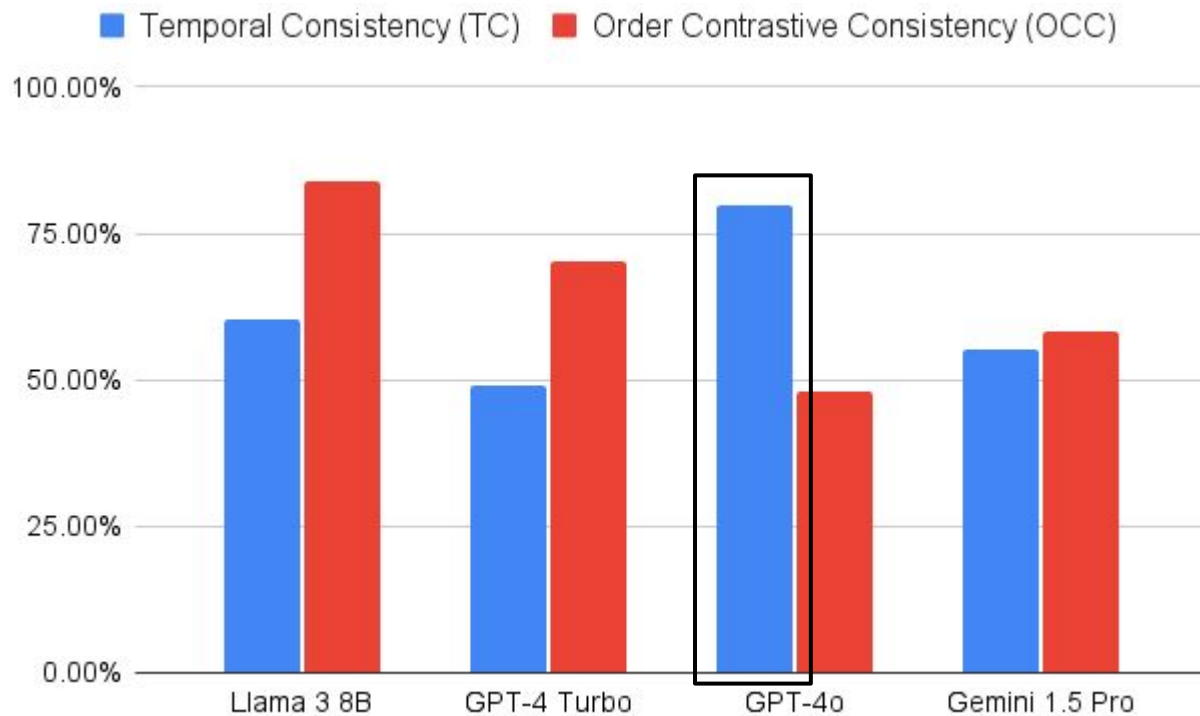
...

Step 12: Whip cream till stiff peaks.

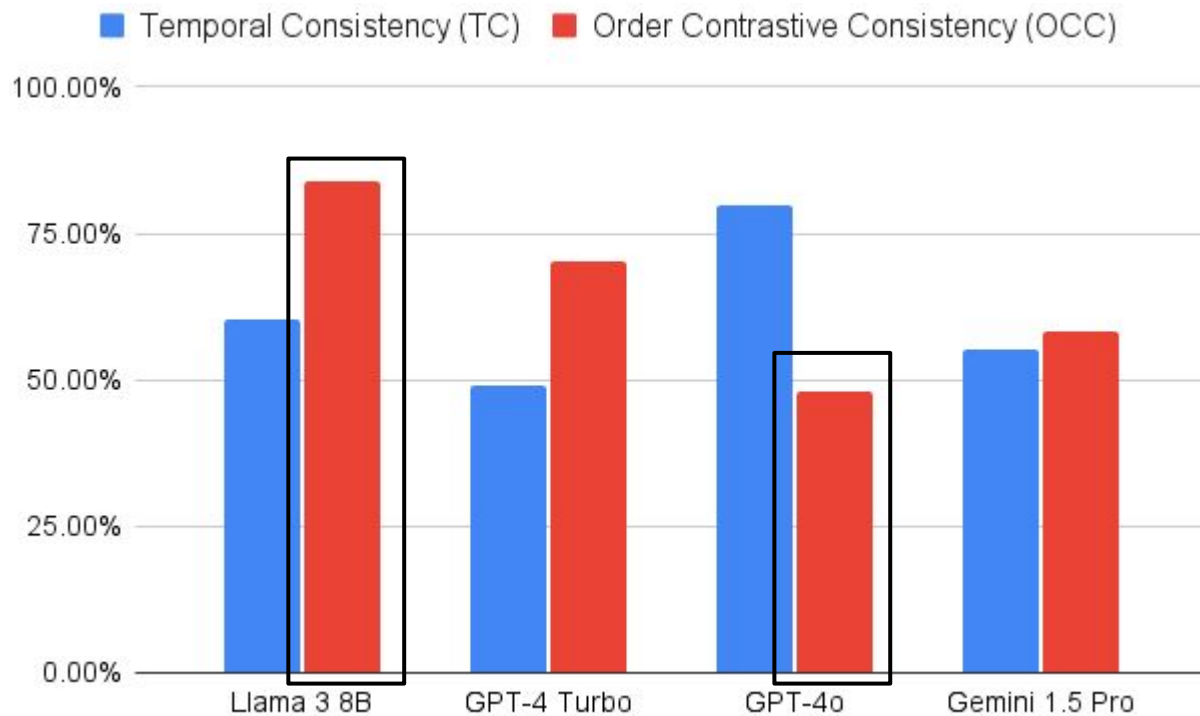
...

The model answer to both questions should be the same.

How Robust are Models?



How Robust are Models?



Conclusion

- We introduce an easy-to-evaluate plan based reasoning benchmark
- SOTA LLMs struggle with this simple task but post-hoc explanations help
- Models are inconsistent and biased towards predicting step dependence



Leaderboard

Contact: ylal@cs.stonybrook.edu