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

Yash Kumar Lal[1*], Vanya Cohen[3*], Nathanael Chambers[2], Niranjan Balasubramanian[1], Raymond J. Mooney[3]

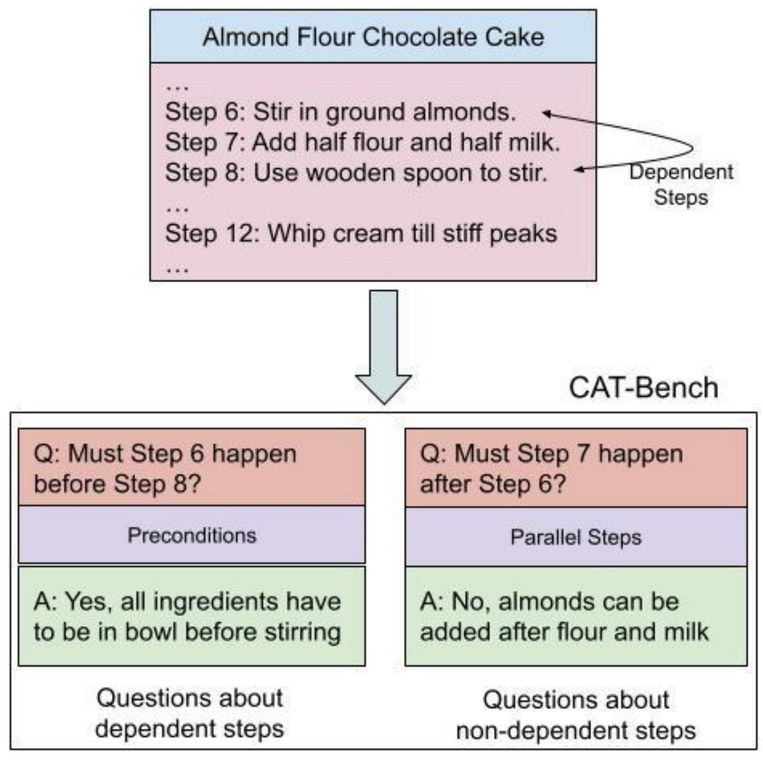
[1]Stony Brook University, [2]US Naval Academy, [3]University of Texas, Austin, *equal contribution

Stony Brook University
Computer Science

MOTIVATION

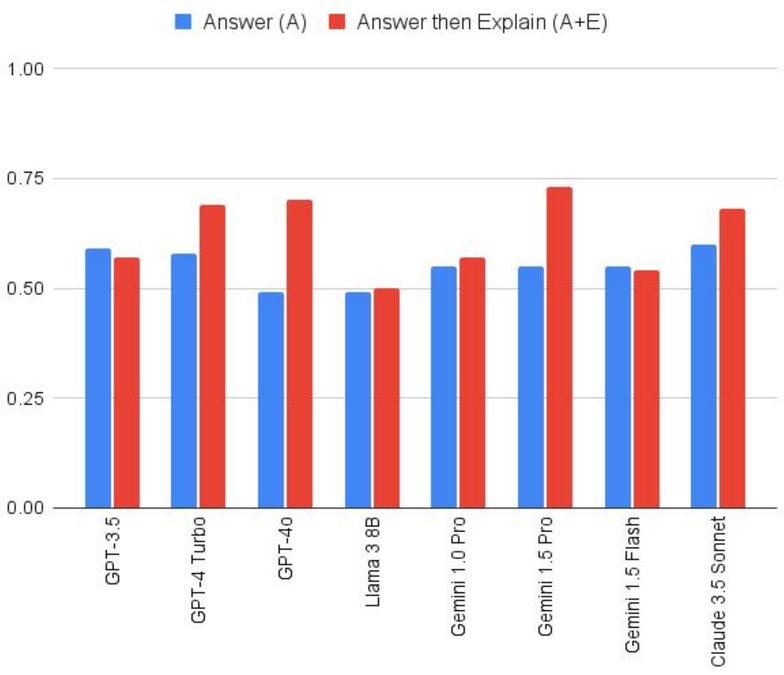
- ➔ We cannot perform plans in the real world to verify whether they are executable and accomplish the goal
- ➔ Simulation worlds are restrictive and do not allow all actions that we can perform in the real world
- ➔ Need for proxy evaluations to test understanding of plans
- ➔ If you understand a piece of text (here, a plan), you should be able to answer all questions about it
- ➔ Holistic **question-driven evaluation** is realistic, safe and cheaper

TASK AND EVALUATION



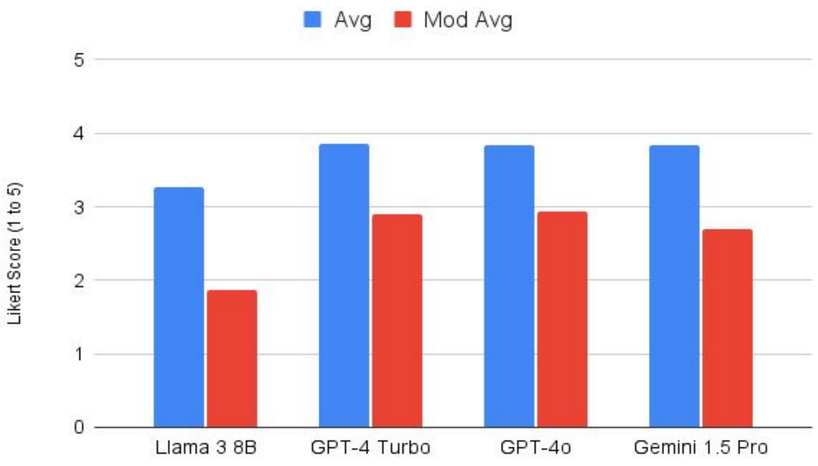
- ➔ **Binary Dependency Prediction**
  - ◆ Must Step 6 happen before Step 8?
  - ◆ Must Step 8 happen after Step 6?
- ➔ F1 Score
  - ◆ Binary dependency prediction
- ➔ Temporal Consistency
  - ◆ Are models consistent in their before/after answers?
- ➔ Order Contrastive Consistency
  - ◆ Are models consistent in their before/after answers?
- ➔ Human Evaluation
  - ◆ Free-form explanations

MODEL BENCHMARKING



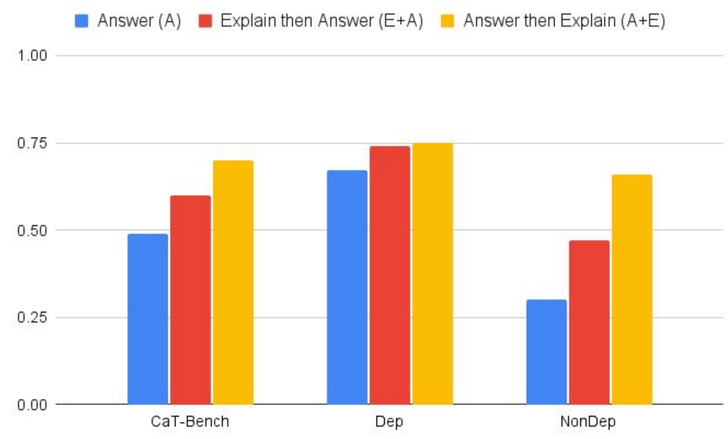
- ➔ Models struggle to understand causal dependencies within natural language plans
- ➔ Often, they are only as good as random chance
- ➔ Models are biased towards predicting causal dependence
- ➔ Prompting them to also provide explanations helps!
- ➔ Explanations also help predict long-range dependencies better

EXPLANATION QUALITY



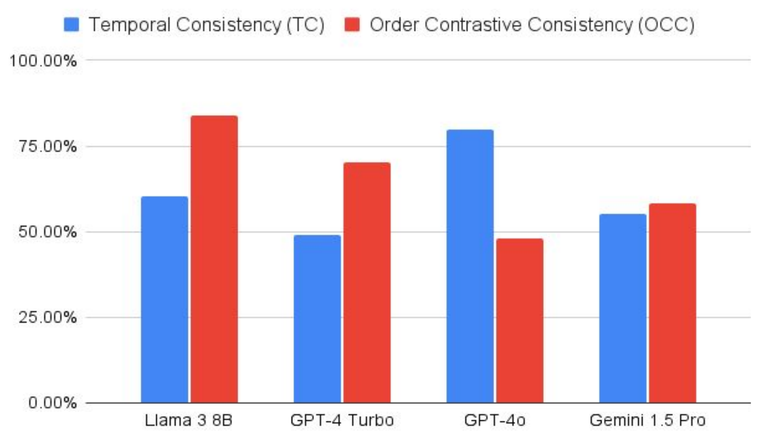
- ➔ Larger models produce more convincing explanations
- ➔ To account for faithfulness to their prediction, we use ModAvg
- ➔ Even the best model scores < 50%
- ➔ Humans **don't** agree with models

REASON OR JUSTIFY?



- ➔ Chain of thought **struggles**!
- ➔ Post-hoc explanations are better than intermediate reasoning
- ➔ Other prompting techniques do not help much

MODEL INCONSISTENCY



- ➔ Models are inconsistent in their reasoning about the same pair of steps (TC)
- ➔ They change predictions for plans with altered step order (OCC)

IMPROVING MODELS

- ➔ Multi-hop dependency: Failure to understand that two steps might be related through an intermediate step
- ➔ Effects: Failure to understand that an effect of the preceding step leads to the succeeding step
- ➔ Preconditions: Failure to understand a condition that needs to be satisfied for a step to happen
- ➔ Irrelevant Answers: Producing explanations that are unrelated to the step being asked about