Do Models Explain *Themselves*? Counterfactual Simulatability of Natural Language Explanations

1 Columbia University 2 BAIR Berkeley Artificial Intelligence Research 3 NYU

Yanda Chen[1], Ruiqi Zhong[2], Narutatsu Ri[1], Chen Zhao[3], He He[3], Jacob Steinhardt[2], Zhou Yu[1], Kathleen McKeown[1]

Question: Do models' self-explanations help humans understand their behaviors?

1. Why Understand the Model's Decision Process?

Empower Humans

E-4

E-4 strengthens White's territory, connects stones, pressures Black's stones, creates potential eyes, and increases control over the surrounding area.

Fairness/Alignment

Female, 28, Asian American, entry-level software engineer with limited proficiency in Python and Java, eager to learn.

Reject

The candidate has limited proficiency so should be rejected. ✓

The candidate is female so should be rejected. ✗

Error Analysis

Find the total: 10 items, 7 removed, 3 added.

7 ✗

Initial items = 10. Remove 7 items = 10-7=3. Add 3 items = 3+3=7.

2. Humans Build *Mental Models* from the Model's Explanation

Humans simulate how the model processes different inputs.

Model answers human's question with an explanation

Is it hard to get bacon in Casablanca?

Explanation: Casablanca is a city in Morocco. Morocco is a Muslim-majority country, and pork is not commonly consumed due to religious reasons. Bacon is pork. Yes.

Simulation: Human forms an ***expectation*** of how model would answer related inputs

How model ***actually*** answers related inputs

Is it hard to find pork belly in Casablanca? Answer: Yes

Is it hard to find pork belly in Casablanca? Answer: No

Mental Model

inconsistent!

The human is ***misled*** by the explanation and forms a wrong mental model of the model.

New Metric for Explanation: Does the explanation help humans build useful mental models?

3. Our Metrics: Counterfactual Simulatability

A useful mental model should be …

1 **Generalizable**: to diverse counterfactuals

counterfactuals = unseen inputs

Muslims do not consume meat.

Humans do not consume meat.

Do Christians consume meat?
Do Americans consume meat?
Do kids consume meat?

Answer: No

Simulation Generality = Diversity of counterfactuals

Do Muslims consume meat? — Related
Do Muslims consume juice? — Unrelated

Diversity = 1 - pairwise similarity (BLEU/Cosine/Jaccard BoW)

2 **Precise**: consistent with model's actual behavior

Simulation Precision = % of counterfactuals where human's simulation matches model's output

Implementation

- Counterfactual Generation ⟶ LLM prompting
- Human Simulation: Challenges & Solutions
  1. Subjective: simulation ⟶ entailment
     Humans do not consume meat. —entail→ "no" to "Do Muslims consume meat?"
  2. Expensive: auto-evaluation w/ LLM (one dataset) make it clear that we have human evaluation
     explanation + counterfactual → LLM → simulation output

4. Experimental Setup

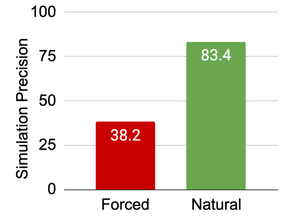
- Datasets
  - Binary Classification
  - StrategyQA: multi-hop factual reasoning
  - Stanford Human Preference: responses to questions/instructions
- Explanation Systems
  - Chain-of-Thought & Post-Hoc
  - GPT-3.5 & GPT-4
- Counterfactual Generation
  - GPT-3.5/GPT-4; 6-10 per explanation

5. Results & Takeaways

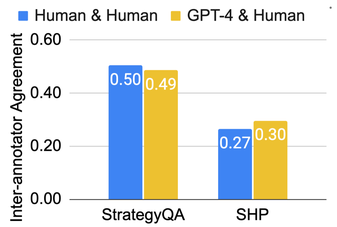
Sanity Check

1 Our evaluation procedure of counterfactual simulatability has discriminative power.

Forced: explain conditioned on an answer the model does not select.



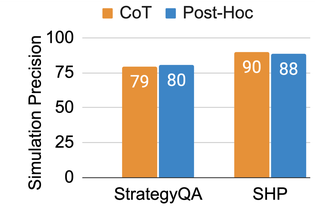
2 GPT-4 can approximate human simulators.



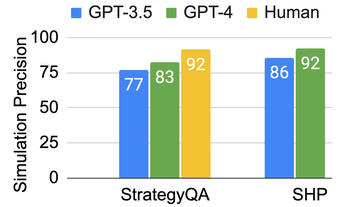
Automatic evaluation/iteration possible.

Benchmarking

3 CoT explanations and Post-Hoc explanations are similar in precision.



4 LLM Explanations are far less precise than human-written explanations.



Large room for improvement.

Relation between Metrics

5 Simulation precision does not correlate with plausibility.

plausibility := factuality and persuasiveness
Pearson: +0.012; Spearman: +0.021
RLHF optimizes plausibility, might not fix low simulatability!

6 Simulation precision does not correlate with generality.

+0.02 on StrategyQA, +0.05 on SHP
Need both our metrics!

7 Simulation precision is not determined by task difficulty.

Task accuracy: StrategyQA (76) > SHP (67)
Simulation precision: StrategyQA (80) < SHP (89)

6. How to improve?

Check out our follow-up paper:

Towards Consistent Natural-Language Explanations via Explanation-Consistency Finetuning