Evaluating LLM-Simulated Conversations in Modeling Inconsistent and Uncollaborative Behaviors in Human Social Interaction

Ryo Kamoi[1†], Ameya Godbole[2†], Binglin Zhou[1], Xiaoxin Lu[1], Longqi Yang[3], Rui Zhang[1], Mengting Wan[3], Pei Zhou[3]

[†]Work done during an internship at Microsoft Corporation

[1] PennState [2] USC University of Southern California [3] Microsoft

Q: Do LLM-Simulated Conversations Reproduce Human Group Behavior?

A. Not Yet. LLM-Simulated Conversations Remain Unreliable Proxies for Human Group Behavior and Are Difficult to Control.

Motivation

LLM-simulated conversations are increasingly being used as proxies for human conversations in applications from social media prototyping to hypothesis testing in social sciences.

What Did We Do?

This is the First Work Focusing on Analyzing Inconsistent and Uncollaborative Behaviors

- Prior work implicitly assumes that human conversations are consistent and collaborative
- We introduce CoCoEval, which provides turn-level detection of 10 types of inconsistent and uncollaborative behaviors
- We compare the frequencies of inconsistent and uncollaborative behaviors of LLMs and humans

Findings

1. **Human conversations exhibit natural complexity due to mistakes**
   - Humans make misunderstandings and interruptions
   - Complexities arise in the correction of mistakes and negotiations
2. **LLM-simulated conversations exhibit different complexities**
   a. Too efficient with vanilla prompting
   b. Too conflicting when prompted to reproduce human complexities
   c. Amplify specific issues like repetition when fine-tuned



CoCoEval: Evaluation Framework with Turn-Level Behavior Detection

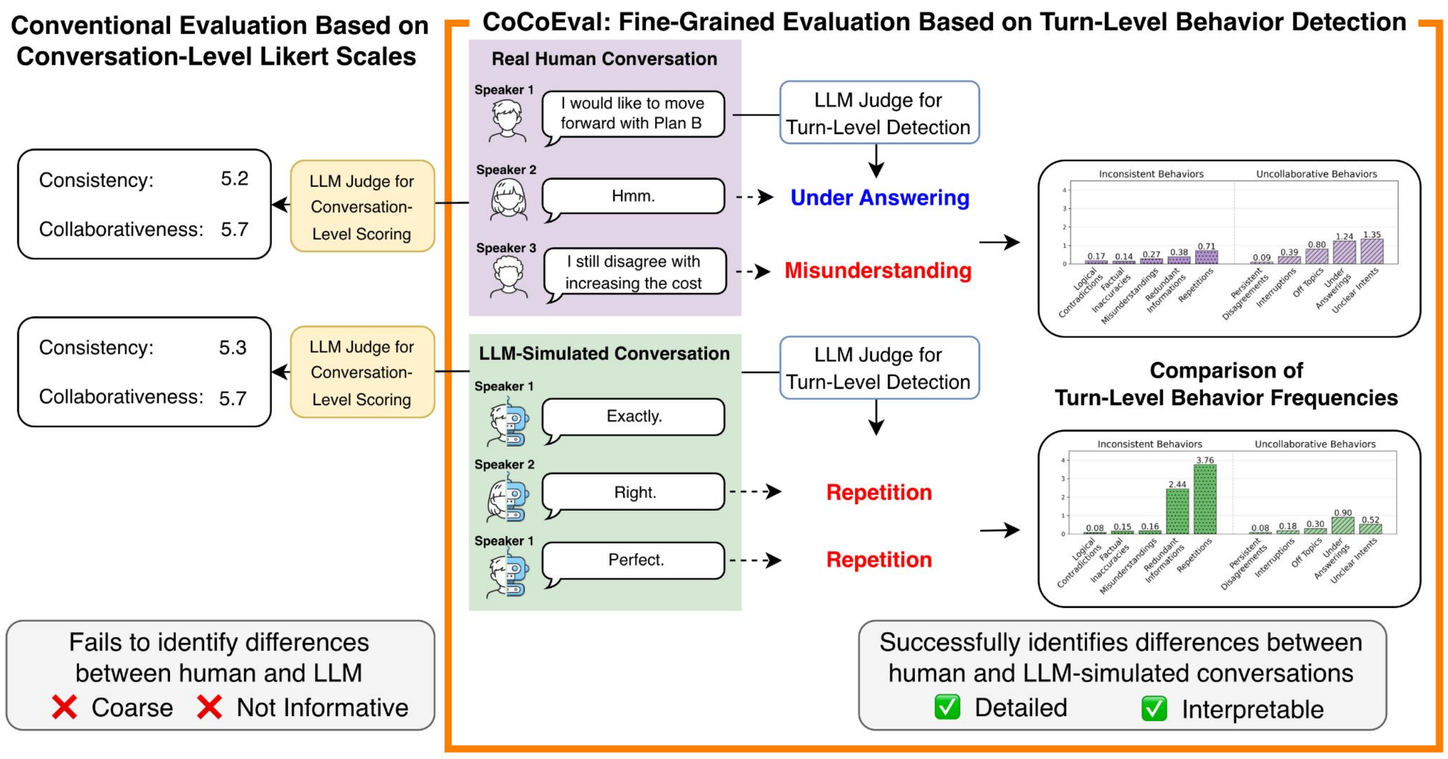
Method: Turn-Level Detection of 10 Behaviors

Inconsistent Behaviors	Uncollaborative Behaviors
Logical Contradiction	Persistent Disagreement
Factual Inaccuracy	Interruption
Misunderstanding	Off Topic
Redundant Information	Under Answering
Repetition	Unclear Intent

New Benchmark for Direct Comparison

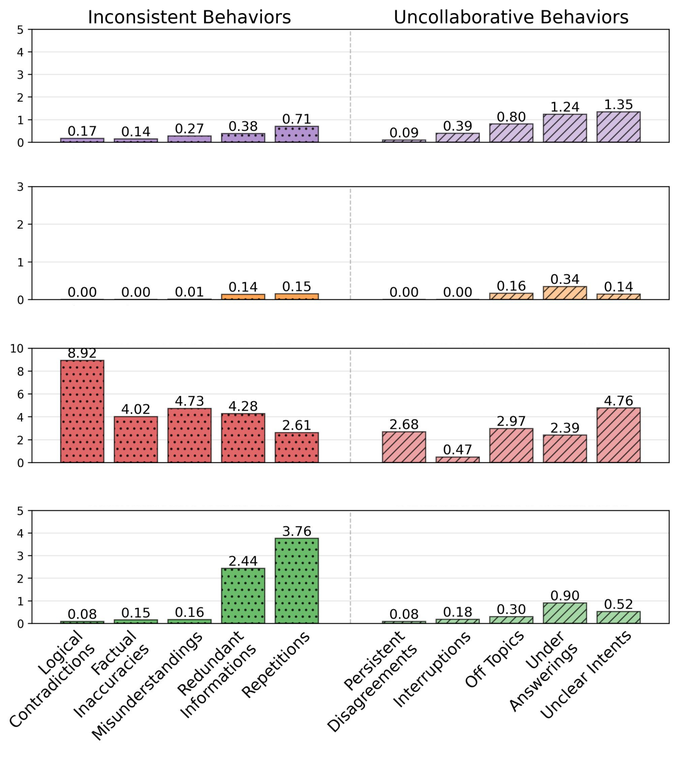
- A benchmark for generating 30-turn continuations conditioned on 30 turns of real human conversation
- Continuations generated from the same context are expected to follow similar behavioral distributions

Domain	Source Human Conversation Resources
Business	QMSum Product
Academic	QMSum Academic
Governmental	QMSum, NCPC
Debates	SIM, IQ2



Results

CoCoEval Provides a Detailed and Interpretable Comparison



Human conversations are not perfect and include inconsistent and uncollaborative behaviors at moderate frequencies

Vanilla prompting results in overly consistent and collaborative conversations

Instructions to reproduce human issues result in overproduction of the behaviors

Supervised fine-tuning on human conversations tends to overproduce specific behaviors like repetition

Conventional Metrics Fail to Separate Human and LLMs

Popular conversation-level Likert scale also captures the observed overall trend, but fails to separate humans and LLMs in some settings.

In contrast, CoCoEval identifies differences in per-behavior frequencies, such as overproduction of repetition and redundancy.

Generation Method	Conventional Conversation-Level Scores	
	Consistency	Collaborative
Human	**5.2**	**5.7**
Vanilla Prompt (GPT-4.1, 30 turns per call)	9.5	9.7
Reproduce Prompt (GPT-5.1, 1 turn per call)	**4.6**	**4.5**
SFT (GPT-4.1, 1 turn per call)	**5.3**	**5.7**