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

Anonymous ACL submission

Abstract

Simulating human conversations using large language models (LLMs) has emerged as a scalable methodology for modeling human social interaction. This paper reconsiders the evaluation of simulated conversations by explicitly recognizing that human conversations inherently involve inconsistent and uncollaborative behaviors, such as misunderstandings and interruptions. Since these behaviors contribute to the complexity of human social interaction, we argue that LLM-simulated conversations should reproduce them with frequencies and patterns comparable to those observed in human conversations. To support a detailed and interpretable evaluation of these behaviors, we introduce CoCoEval, a framework consisting of an evaluation scheme based on turn-level detection of 10 types of inconsistent and uncollaborative behaviors and a benchmark for simulating conversations in professional scenarios involving collaboration and conflict. Using CoCoEval, we compare human conversations with those simulated by GPT-4.1, GPT-5.1, and Claude Opus 4. The results show that (1) LLM-simulated conversations exhibit far fewer inconsistent and uncollaborative behaviors than human conversations under vanilla prompting, and (2) prompt engineering and supervised fine-tuning do not provide reliable control over these behaviors, often leading to the overproduction of specific behaviors. CoCoEval identifies gaps between human and LLM-simulated conversations that are not captured by conventional evaluation based on conversation-level Likert scales, raising caution regarding the use of LLMs as proxies for human social interaction. The benchmark, model outputs, and prompts are in the supplementary material.

1 Introduction

Simulating human conversations by generating utterances of all participants using large language models (LLMs) has recently emerged as a scalable methodology for modeling human social interaction, with potential applications in prototyping social platforms such as social media (Park et al., 2023), testing hypotheses in social science (Manning et al., 2024), and helping decision-makers forecast how people may respond to new policies (Sreedhar and Chilton, 2024). However, it remains an open question to what extent LLMs can reproduce human behaviors and serve as a reliable proxy for human social interaction.

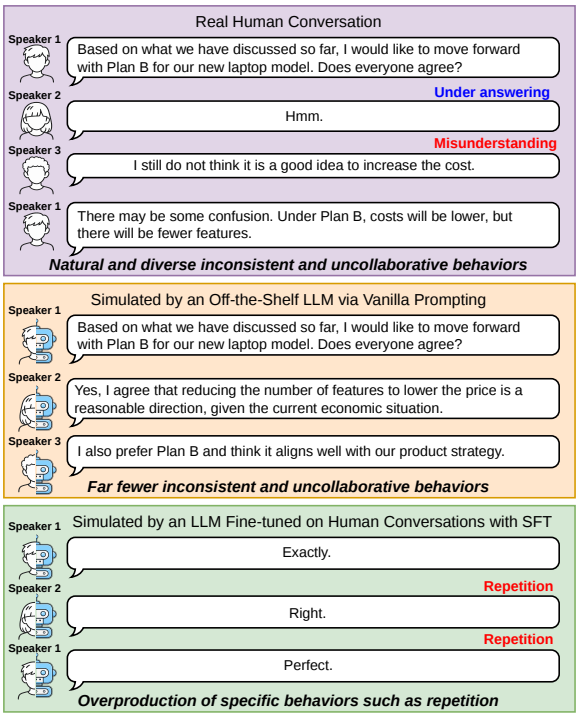


Figure 1: Human conversations inherently involve inconsistent and uncollaborative behaviors, which constitute the complexity of human social interaction (top). CoCoEval reveals that LLM-simulated conversations fail to reproduce these behaviors, often containing far fewer of them (middle) or overproducing specific behaviors (bottom), depending on the simulation setup.

This paper reconsiders the evaluation of LLM-simulated conversations by explicitly recognizing

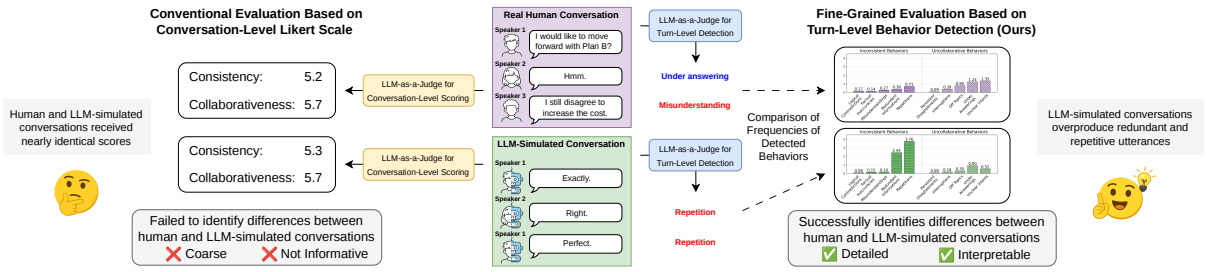


Figure 2: CoCoEval introduces a fine-grained evaluation scheme based on turn-level detection of inconsistent and uncollaborative behaviors (right). Compared to conventional evaluation based on conversation-level Likert scales (left), our method provides more interpretable and detailed results, better capturing property differences between humans and LLMs. This figure shows an example for GPT-4.1 fine-tuned on human conversations via SFT.

that human conversations are not perfectly consistent or collaborative and inherently involve inconsistent behaviors, such as misunderstandings and repetition, as well as uncollaborative behaviors, such as interruptions and disagreements (Grice, 1975; Schegloff, 1992; Pickering and Garrod, 2004). These behaviors contribute to the complexity of human social interactions by giving rise to downstream social behaviors such as recovering from misunderstandings or persuading other participants. Therefore, we argue that reproducing inconsistent and uncollaborative behaviors is necessary for modeling human social interaction using LLM-simulated conversations, and this work analyzes whether LLMs can reproduce these behaviors with frequencies and patterns comparable to those in human conversations. This work addresses two research gaps. (1) First, prior work often implicitly assumes that higher consistency and collaborativeness correspond to better simulation quality (Zhou et al., 2024b; Wang et al., 2025). This assumption is misaligned with the goal of simulating human social interaction, which is not perfectly consistent or collaborative. (2) Second, prior work typically evaluates simulated conversations using conversation-level Likert scales (Park et al., 2023; Zhou et al., 2024b), such as consistency and collaborativeness rated on a 1-10 scale, which does not provide a detailed or interpretable analysis.

We introduce CoCoEval, an evaluation framework comprising an evaluation scheme and a benchmark designed to facilitate detailed and interpretable analysis of inconsistent and uncollaborative behaviors in human and LLM-simulated conversations. First, we propose a fine-grained evaluation scheme that detects inconsistent and uncollaborative behaviors at the turn level. We define 10 types of behaviors based on taxonomies for dialogue generation (Higashinaka et al., 2021; Wang et al., 2023), as in Table 1. A successful simulation should include these behaviors with frequencies and patterns comparable to those in human conversations. We also show the reliability of LLM-as-a-Judge in CoCoEval through agreement with human evaluation and employ it for scalable evaluation. Second, we introduce CoCoEvalBench, a benchmark to evaluate LLMs on simulating 30-turn conversation continuations given conversational metadata and history. It is created from human conversations in collaborative or conflicting situations in QMSum (Zhong et al., 2021), NCPC (NCPC, 2024), SIM (Fan et al., 2024), and IQ2 (Zhang et al., 2016), spanning business, academic, and governmental meetings as well as debates.

Using CoCoEval, we evaluate GPT-4.1, GPT-5.1, and Claude Opus 4 in widely used conversation simulation setups. CoCoEval reveals discrepancies between human and LLM-simulated conversations in inconsistent and uncollaborative behaviors, as in Figure 1. First, under vanilla prompting, LLMs generate conversations with far fewer inconsistent and uncollaborative behaviors than human conversations (§5.2). Second, prompt engineering does not provide reliable control over these behaviors. LLMs overproduce these behaviors when explicitly instructed to include them at frequencies observed in human conversations (§5.3). Third, supervised fine-tuning on human conversations results in overproducing a narrow range of behaviors such as repetition and redundancy (§5.4).

In summary, our analysis of inconsistent and uncollaborative behaviors in LLM-simulated conversations using CoCoEval shows that the properties of simulated conversations are highly sensitive to simulation settings, and that none of the evaluated settings reproduces the frequencies and patterns of these behaviors observed in human conversations. The results also demonstrate an advantage of Co-

	Name	Definition	Example
Inconsistent Behaviors	Logical Contradictions	Utterances that contradict earlier statements by the same speaker or the speaker's metadata.	One participant says "The remote control runs on a single AA battery," and later says "It requires two AA batteries."
	Factual Inaccuracies	Utterances that contain factually incorrect statements.	A participant claims that infrared signals can pass through walls.
	Misunder-standings	Utterances that reflect a misinterpretation of information provided by other participants.	A participant states that the remote control uses only a touch screen, but another later assumes that it has physical buttons.
	Redundant Information	Utterances that contain unnecessarily lengthy content and add little new informational value.	A participant gives a long explanation about previous projects without adding new information.
	Repetition	Utterances that unnecessarily repeat information stated earlier in the conversation.	One participant says "I will bring the documents," and another asks "So you will bring the documents, right?"
Uncollaborative Behaviors	Persistent Disagreement	Utterances that repeatedly reject or others' positions despite established consensus or supporting evidence.	A participant continues to reject a proposed plan even after others agree and provide supporting evidence for it.
	Interruptions	Utterances that disrupt the conversation by cutting off another participant.	A participant interrupts another speaker mid-explanation and leaves the utterance unfinished, like "... the key point is".
	Off-topic Responses	Utterances that deviate substantially from the current conversational topic.	During the middle of a project meeting discussion, a participant shares unrelated personal stories.
	Under Answering	Utterances that give evasive responses and fail to address preceding questions.	When asked "Will this policy increase taxes?", a participant replies "We are focused on supporting working families."
	Unclear Intent	Utterances that are ambiguous or fail to clearly convey the speaker's intentions.	When asked whether to agree or disagree, a participant responds only with "Maybe."

Table 1: Definitions of inconsistent and uncollaborative behaviors used in the fine-grained evaluation in CoCoEval.

CoEval's evaluation scheme based on turn-level behavior detection, which identifies differences between human and LLM-simulated conversations that conventional conversation-level Likert scale evaluation fails to capture (§6), as illustrated in Figure 2. Below, we summarize our contributions:

- We are the first to analyze LLM-simulated conversations with a focus on inconsistent and uncollaborative behaviors, which are inherent in human conversations and should be reproduced to simulate human social interaction.
- We introduce CoCoEval, a framework for analyzing inconsistent and uncollaborative behaviors in human and simulated conversations. It consists of an evaluation scheme designed to provide detailed and interpretable analysis, along with a benchmark featuring professional scenarios.
- Our analysis using CoCoEval reveals that LLM-simulated conversations often exhibit inconsistent and uncollaborative behaviors with frequencies and patterns that differ from those in human conversations, raising concerns about their use as proxies for human social interaction.

2 Related Work

We study the simulation of multi-party conversations in which LLMs simulate all participants (Park et al., 2023; Manning et al., 2024; Sreedhar and Chilton, 2024; Zhou et al., 2024b). This setting has been studied for modeling human social interaction beyond simulating a single individual, in contrast to situations where LLMs simulate a single participant (Elhilali et al., 2025) or human-LLM conversations (Zheng et al., 2023; Wu et al., 2025).

Consistency and collaborativeness in this paper cover most commonly used evaluation dimensions in conversation simulation, including coherence with conversation history (Maharana et al., 2024; Chu et al., 2024), alignment with persona and role (Frisch and Giulianelli, 2024; Abdulhai et al., 2025), adherence to social norms (Li et al., 2023), and social interaction capabilities (Zhou et al., 2024b; Piedrahita et al., 2025), while a few dimensions, such as behavioral diversity (Pang et al., 2020), are not covered. Evaluation in prior work commonly relies on conversation-level Likert scales, whereas we introduce an evaluation scheme based on turn-level behavior detection to enable more detailed and interpretable evaluation.

3 CoCoEval

We introduce CoCoEval, an evaluation framework for analyzing LLM-simulated conversations with a focus on inconsistent and uncollaborative behaviors, which are inherent to human conversations (Schegloff, 1992; Grice, 1975; Pickering and Garrod, 2004) and contribute to the complexity of human social interaction. CoCoEval comprises an evaluation scheme (§3.1) and a benchmark (§3.2) designed to facilitate detailed and interpretable comparison between human- and LLM-

generated conversations with respect to these behaviors. We implement the evaluation using LLM-as-a-Judge to support scalable evaluation (§3.3).

The definitions of consistency and collaborativeness in CoCoEval follow prior discussions in conversation evaluation. **Consistency** includes coherence with previous utterances, adherence to the specified persona, and correctness with respect to world knowledge (Park et al., 2023; Frisch and Giulianelli, 2024; Abdulhai et al., 2025), as well as the avoidance of unnecessary redundancy and repetition (Higashinaka et al., 2021; Liu et al., 2025). **Collaborativeness** is the extent to which a speaker's contributions support shared objectives and facilitate the goals of other participants (Zhou et al., 2024b; Piatti et al., 2024; Piedrahita et al., 2025; Nonomura and Mori, 2025), without delaying or impeding progress and agreement.

3.1 Evaluation Scheme

CoCoEval adopts two evaluation schemes. The first is the overall evaluation based on Likert scale scores, which is a common evaluation approach in conversation simulation (Zhou et al., 2024b; Wang et al., 2025). However, it is coarse-grained and unsuitable for detailed analysis. To provide a more detailed and interpretable evaluation, we introduce the fine-grained evaluation that detects inconsistent and uncollaborative behaviors at the turn level.

Overall evaluation assigns two conversation-level Likert scale scores, one for consistency and one for collaborativeness, each on a 1-10 scale. We provide definitions and scoring rubrics describing the criteria for each score range (Table 19).

Fine-grained evaluation detects 10 types of inconsistent and uncollaborative behaviors at the turn level. We derive 10 types of inconsistent and uncollaborative behaviors shown in Table 1 by adapting nine types of behaviors from evaluation taxonomies in dialogue generation (Higashinaka et al., 2021; Mahajan et al., 2022), and we added "persistent disagreement" that reflects collaborativeness in negotiation and debates (Wang et al., 2023; Liu et al., 2025). For each category, we detect turns that exhibit the corresponding behaviors and count the number of occurrences in each conversation.

3.2 CoCoEvalBench

We introduce CoCoEvalBench, a benchmark for evaluating LLMs on the conversation continuation task, in which models simulate continuations

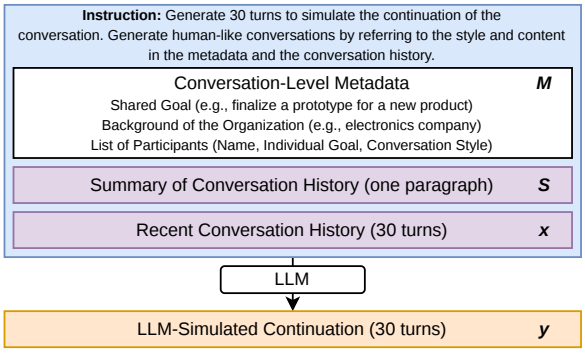


Figure 3: Evaluation task in CoCoEvalBench.

given conversational metadata and history (Figure 3). It covers real-world collaborative or conflicting scenarios spanning business, academic, and governmental meetings, as well as debates. Compared to prior work (Zhou et al., 2024b; Bianchi et al., 2024), we focus on longer conversations with complex social interactions in professional scenarios. Accordingly, we design the task to require simulating 30-turn continuations, which are longer than tasks in most prior work.

Evaluation task. The input consists of metadata M containing the conversation goals and participant information, the 30 preceding turns $x = [x_{-30}, \ldots, x_{-1}]$, and a paragraph-length summary S of the conversation prior to x. Each turn in x is represented as a speaker-utterance pair $x_i = (p_i, u_i)$. Given this input, LLMs are instructed to generate 30 continuation turns $y = [y_1, \ldots, y_{30}] \sim \text{LLM}(M, S, x)$, where each turn is a pair of a speaker and an utterance. Our benchmark includes a human-generated reference continuation $r = [r_1, \ldots, r_{30}]$. We adapt this setup to support direct comparison between LLMs and humans. Since the simulated continuation y and the human reference continuation r share the same context (M, S, x), they are expected to reflect similar frequencies and patterns of inconsistent and uncollaborative behaviors that are coherent with the metadata and history, and substantial deviations in these behaviors indicate simulation failures.

Benchmark creation. We create CoCoEvalBench from real-world human conversations in which multiple participants pursue shared or conflicting goals in professional scenarios involving complex human social interaction. Specifically, we use six public conversation datasets: business meetings in QMSum Product, academic meetings in QMSum Academic, governmental meetings

Category	Source Dataset	# Data Points	Average # Participants	Average # Tokens/Turn
Evaluation Set				
Business	QMSum Product	50	4.0	8.0
Academic	QMSum Academic	50	5.7	10.0
Governmental	QMSum Committee	50	18.5	66.3
	NCPC	50	25.8	50.4
Debates	SIM	50	4.0	16.2
	IQ2	50	10.1	41.3
Training Set				
Business	QMSum Product	500	4.0	9.3
Governmental	QMSum Committee	300	18.6	73.3
Debates	SIM	300	4.0	16.4
	IQ2	350	10.7	47.3

Table 2: Statistics of CoCoEvalBench.

Category	Inter-human Randolph's κ	Human vs. o4-mini Spearman's ρ	# Detected Turns	
			Human	o4-mini
Logical Contradiction	0.804	0.026	6.0	4
Factual Inaccuracy	0.893	0.237	2.4	2
Misunderstanding	0.775	0.254	4.6	6
Redundant Information	0.731	0.185	10.4	12
Repetition	0.592	0.628*	18.4	23
Persistent Disagreement	0.931	0.393*	4.4	2
Interruptions	0.680	0.672*	33.4	12
Off Topic	0.774	0.481*	11.6	15
Under Answering	0.718	0.678*	16.8	38
Unclear Intent	0.760	0.712*	9.8	42

Table 3: Agreement between human evaluation (average of 5 annotators) and LLM-as-a-Judge (o4-mini), as well as inter-human agreement, in the number of detected behaviors for CoCoEval's fine-grained evaluation across 1,080 turns in 36 conversations. *: $p < 0.05$ for ρ.

in QMSum Committee (Zhong et al., 2021) and NCPC (NCPC, 2024), and debates in SIM (Fan et al., 2024) and IQ2 (Zhang et al., 2016). To create the conversation continuation task, we first use o4-mini to extract information from conversations in the source datasets to generate metadata M. Next, we randomly select a start point for the conversation continuation task and extract the preceding 30 turns as history x and the following 30 turns as reference continuation r. We then generate a paragraph-length summary S of the conversation preceding the 30-turn history x using o4-mini.

Statistics. Table 2 shows the statistics of CoCoEvalBench. The evaluation set includes 300 conversations from six datasets, and the training set consists of 1,450 conversations from four datasets.

3.3 LLM-as-a-Judge

To support scalable evaluation, we implement the evaluation schemes using LLM-as-a-Judge. To validate its use as an automated proxy for human evaluation, we collect evaluations from five annotators on human and LLM-simulated conversations for randomly selected cases from CoCoEvalBench and assess their agreement with LLM-as-a-Judge. We use OpenAI o4-mini (OpenAI, 2025c) as LLM-as-a-Judge in Section 5 because of its reasonable agreement with human evaluation. Appendix B provides detailed settings and results using GPT-5.1 Instant and High (OpenAI, 2025a) and Claude Opus 4 (Anthropic, 2025) as LLM-as-a-Judge.

Overall evaluation. The five annotators evaluate the consistency and collaborativeness of 54 conversations generated by humans and GPT-4.1 on a 1-10 scale. The Spearman's rank correlation coefficients between the average human scores and LLM-as-a-Judge (o4-mini) are 0.558 for consistency and 0.614 for collaborativeness. The inter-human agreement, measured by Randolph's Kappa, is 0.284 and 0.280, respectively. We attribute this relatively low inter-human agreement to the subjectivity of the evaluation task. However, LLM-as-a-Judge shows reasonably high agreement with the average human scores, demonstrating its reliability.

Fine-grained evaluation. The five annotators detect 10 types of inconsistent and uncollaborative behaviors at the turn level across 1,080 turns from 36 conversations generated by humans and GPT-4.1. Table 2 reports Spearman's rank correlations between the average human evaluation and LLM-as-a-Judge (o4-mini), as well as inter-human agreement measured by Randolph's Kappa, for the number of detected behaviors in each conversation. The results first show high inter-human agreement, indicating that human annotators can relatively objectively evaluate conversations with this evaluation scheme. The results also show high correlations between human evaluation and LLM-as-a-Judge (o4-mini) for most behavior types that occur more than 10 times in the evaluated 1,080 turns. Specifically, LLM-as-a-Judge (o4-mini) shows statistically significant correlations in detecting all types of uncollaborative behaviors and Repetition, while it shows positive but relatively small correlations for the remaining four inconsistent behaviors. These inconsistent behaviors, except for Redundant Information, are rare in the evaluated domains and therefore do not substantially affect our analysis, although further analysis is needed when applying the framework to domains with more frequent inconsistent behaviors. The table also presents the

total number of detected behaviors in 1,080 turns, showing that LLM-as-a-Judge (o4-mini) largely captures the overall trends observed in human evaluation, although it tends to overdetect Underanswering and Unclear Intent.

4 Conversation Simulation Setup

We evaluate widely used conversation simulation settings: off-the-shelf LLMs with prompt engineering and LLMs fine-tuned on human conversations. We also study how simulation parameters influence inconsistent and uncollaborative behaviors.

LLMs. We evaluate conversation simulated by GPT-4.1 (OpenAI, 2025b), GPT-5.1 Instant (OpenAI, 2025a), and Claude Opus 4 (Anthropic, 2025).

Prompt. We evaluate two types of instructions to assess the influence of prompt engineering. First, in **vanilla prompting** (Table 16), we instruct LLMs to generate *"human-like conversations by referring to the conversation history"* and *"mimic the communication style and content in the metadata and the conversation history."* Second, we introduce **taxonomy-guided prompting** (Table 18) that explicitly instructs LLMs to incorporate inconsistent and uncollaborative behaviors with frequencies and patterns comparable to those observed in the conversation history, and provide the list of the 10 behavior types in Table 1 as follows:

> *Human conversations may involve inconsistent and uncollaborative behaviors [...] The following list outlines representative inconsistent and uncollaborative behaviors: [...]*

Generating multiple turns per LLM call. Generating multiple consecutive turns in a single LLM call is computationally efficient, but generating one turn at a time better reflects the turn-by-turn nature of human conversations (Zhou et al., 2024a). To examine the impact of these generation strategies, we evaluate settings that generate 1, 5, and 30 turns per call, as illustrated in Figure 5. For Claude Opus 4, we only evaluate the settings generating 1 or 5 turns per call, as it frequently reaches its maximum output length when generating 30 turns per call.

SFT on human conversations. Supervised fine-tuning is a popular approach for aligning LLMs for conversation simulation (Wang et al., 2025; Chuang et al., 2025). We fine-tune GPT-4.1 on human conversations from the training set of CoCoEvalBench. We fine-tune three models that generate 1, 5, and 30 turns per call, respectively.

		# Turns Per Call	Consistency	Collaborativeness
Human		–	5.2 [4.8, 5.6]	5.7 [5.3, 6.0]
GPT-4.1	Vanilla Prompting	1	8.2 [7.9, 8.5]	8.6 [8.3, 8.8]
		5	9.6 [9.4, 9.7]	9.6 [9.5, 9.7]
		30	9.5 [9.4, 9.7]	9.7 [9.6, 9.8]
	Taxonomy-Guided Prompting	1	1.3 [1.2, 1.4]	1.3 [1.2, 1.3]
		5	2.8 [2.5, 3.0]	2.9 [2.6, 3.1]
		30	9.3 [9.1, 9.5]	9.4 [9.3, 9.6]
	Supervised Fine-tuning	1	5.3 [4.9, 5.7]	5.7 [5.4, 6.1]
		5	6.6 [6.3, 7.0]	7.2 [6.8, 7.5]
		30	6.3 [5.9, 6.7]	6.7 [6.3, 7.0]
GPT-5.1	Vanilla Prompting	1	9.2 [9.0, 9.4]	9.2 [8.9, 9.4]
		5	9.7 [9.6, 9.8]	9.8 [9.7, 9.9]
		30	9.8 [9.7, 9.9]	9.9 [9.8, 9.9]
	Taxonomy-Guided Prompting	1	4.6 [4.2, 5.0]	4.5 [4.2, 4.9]
		5	6.5 [6.1, 6.9]	6.5 [6.2, 6.9]
		30	9.3 [9.1, 9.5]	9.3 [9.1, 9.5]
Claude Opus 4	Vanilla Prompting	1	9.8 [9.7, 9.9]	9.9 [9.8, 9.9]
		5	9.8 [9.7, 9.9]	9.9 [9.8, 9.9]
	Taxonomy-Guided Prompting	1	1.5 [1.4, 1.6]	1.4 [1.3, 1.5]
		5	1.7 [1.6, 1.9]	1.6 [1.5, 1.7]

Table 4: Results of CoCoEval's overall evaluation using conversation-level scores on a 1-10 scale, evaluated by o4-mini. Each setup is evaluated on 300 simulated conversations for CoCoEvalBench. The intervals represent 95% bootstrap confidence intervals.

5 Results

We compare human and LLM-simulated conversations using CoCoEval on the conversation continuation task in CoCoEvalBench, evaluated with LLM-as-a-Judge (o4-mini). Our analysis addresses the following research questions.

RQ1: How much do **human conversations** exhibit inconsistent and uncollaborative behaviors? (§5.1)

RQ2: With **vanilla prompting**, how well off-the-shelf LLMs simulate inconsistent and uncollaborative behaviors in human conversations? (§5.2)

RQ3: Can **prompt engineering** control inconsistent and uncollaborative behaviors in conversations simulated by off-the-shelf LLMs? (§5.3)

RQ4: Can **supervised fine-tuning** enable LLMs to reproduce inconsistent and uncollaborative behaviors in human conversations? (§5.4)

RQ5: Does **generating multiple turns per LLM call** affect inconsistent and uncollaborative behaviors in LLM-simulated conversations? (§5.5)

5.1 Human Conversations

The overall evaluation results in Table 4 show consistency and collaborativeness scores of 5.2 and 5.7 out of 10 for human conversations, indicating that they are not perfectly consistent or collabo-

rative. The fine-grained evaluation results in Figure 4a provide a more detailed view, showing that human conversations contain all 10 evaluated behavior types at moderate frequencies, with Unclear Intents, Under Answering, Off Topics, and Repetitions occurring relatively frequently.

Answer to RQ1 about Human Conversations

Human conversations involve diverse types of inconsistent and uncollaborative behaviors.

5.2 Vanilla Prompting

The overall evaluation in Table 4 indicates that conversations generated with vanilla prompting exhibit substantially higher consistency and collaborativeness than human conversations. The fine-grained evaluation in Figures 4b and 4c further shows that inconsistent and uncollaborative behaviors are nearly absent, particularly when 30 turns are generated per call, highlighting substantial differences from human conversations (Figure 4a).

Answer to RQ2 about Vanilla Prompting

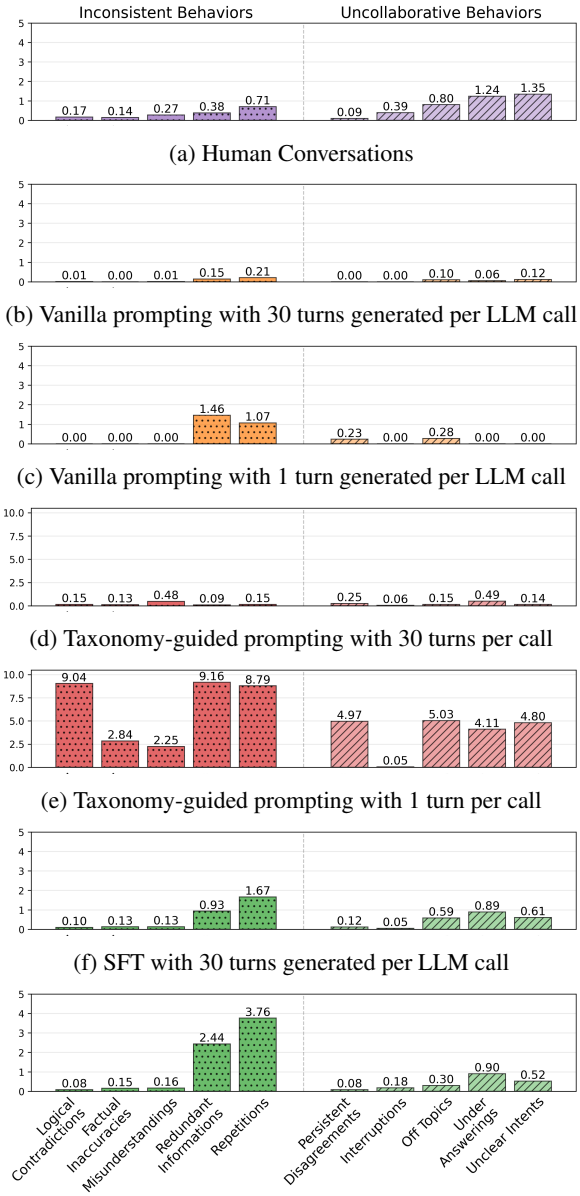
With vanilla prompting, LLM-simulated conversations exhibit far fewer inconsistent and uncollaborative behaviors than human conversations.

5.3 Prompt Engineering

The overall evaluation in Table 4 shows that the taxonomy-guided prompting largely reduces overall consistency and collaborativeness relative to vanilla prompting. The effect is stronger when fewer turns are generated within a single LLM call. The fine-grained evaluation in Figure 4e shows that this prompt substantially increases the frequency of almost all types of inconsistent and uncollaborative behaviors when generating a single turn per call, although this prompt does not instruct to increase these behaviors but rather to reproduce their frequencies in human conversations. The resulting frequencies and patterns differ substantially from those observed in human conversations (Figure 4a).

Answer to RQ3 about Prompt Engineering

Prompt engineering does not provide reliable control over inconsistent and uncollaborative behaviors and can lead to their overproduction when explicit instructions are provided.



(a) Human Conversations

(b) Vanilla prompting with 30 turns generated per LLM call

(c) Vanilla prompting with 1 turn generated per LLM call

(d) Taxonomy-guided prompting with 30 turns per call

(e) Taxonomy-guided prompting with 1 turn per call

(f) SFT with 30 turns generated per LLM call

(g) SFT with 1 turn generated per LLM call

Figure 4: Results of CoCoEval's fine-grained evaluation showing the number of turns exhibiting each behavior type per 30-turn conversation continuation generated by humans and GPT-4.1, averaged over 300 conversations in CoCoEvalBench. The y-axes for taxonomy-guided prompting use different scales.

5.4 SFT on Human Conversations

The overall evaluation in Table 4 shows that GPT-4.1 fine-tuned on human conversations with SFT exhibits levels of consistency and collaborativeness comparable to those of human conversations, especially when generating one turn per call. However, the fine-grained evaluation in Figure 4g shows that the frequencies and patterns of inconsistent and uncollaborative behaviors do not match those of human conversations (Figure 4a), with fine-tuned

GPT-4.1 overproducing Repetition and Redundant Information. We attempt to alleviate this issue using a presence penalty (Appendix C.2), but we observe that the bias toward repetition and redundancy is strong and cannot be easily mitigated. These results indicate that even when fine-tuned on human conversations, LLM-simulated conversations still deviate from human conversations.

Answer to RQ4 about SFT

SFT on human conversations can lead LLMs to overproduce a narrow subset of inconsistent behaviors, such as repetitions and redundancy, remaining misaligned with human conversations.

5.5 Generating Multiple Turns Per LLM Call

The results in Table 4 and Figure 4 consistently show that generating multiple turns per LLM call, particularly 30 turns, introduces a strong bias toward reducing inconsistent and uncollaborative behaviors. In particular, Figure 4d shows that generating 30 turns in a single LLM call almost eliminates the effect of taxonomy-guided prompting, whereas it substantially increases these behaviors when one turn is generated per call. However, generating fewer turns per call still fails to reproduce the distribution of these behaviors observed in human conversations (Figure 4a) and instead tends to over-represent repetitive and redundant utterances, even with the vanilla prompting, as shown in Figure 4c.

Answer to RQ5 about Generation Strategy

The number of turns generated in a single LLM call strongly influences inconsistent and uncollaborative behaviors in simulated conversations, but adjusting this parameter alone does not fully account for the gap from human conversations.

6 Benefits of CoCoEval's Evaluation

The results in Section 5 show that CoCoEval's fine-grained evaluation provides a detailed and interpretable evaluation, and as a result, it captures differences between human and LLM-simulated conversations that are not captured by conventional conversation-level Likert-scale evaluation (the overall evaluation). First, the overall evaluation in Table 4 assigns similar scores to human conversations and LLM-generated conversations for fine-tuned GPT-4.1 when generating one turn per call. However, our manual analysis shows that conversations simulated by fine-tuned GPT-4.1 often exhibit unnatural repetition, as shown in an example in Table 5, where the same speaker repeatedly expresses agreement with prior utterances. In contrast, the fine-grained evaluation in Figure 4g successfully identifies that fine-tuned GPT-4.1 overproduces repetition and redundancy compared to human conversations (Figure 4a). In addition, GPT-5.1 with taxonomy-guided prompting (generating one turn per call) achieves scores similar to those of human conversations in the overall evaluation in Table 4, while our manual analysis shows that these conversations often contain unnaturally frequent misunderstandings, as in an example in Table 6. The fine-grained evaluation in Figure 7f successfully captures this difference, revealing a substantially different distribution of turn-level behaviors, particularly a high frequency of misunderstanding.

Manager	Exactly.
Interface	Yeah, that makes sense.
Designer	That's right.
Interface	Alright.
Manager	Okay.

Table 5: Conversation generated by GPT-4.1 fine-tuned on human conversations for QMSum Product, exhibiting unnatural repetitive utterances that are not human-like.

PhD C	Mm-hmm, so you're saying we'd basically scale the acoustic posteriors with the VAD silence likelihood, right? [...]
Prof. B	No, I wasn't really talking about changing the posteriors [...]
PhD A	So we'd be tweaking the silence posteriors a bit? [...]
Prof. B	No, you wouldn't be touching the posteriors [...]

Table 6: Conversation generated by GPT-5.1 with taxonomy-guided prompting for QMSum Academic, exhibiting unnaturally repetitive misunderstanding.

7 Conclusion

This is the first work to conduct a detailed analysis of LLM-simulated conversations with a focus on inconsistent and uncollaborative behaviors, which are inherent to human conversations and contribute to the complexity of human social interactions. We introduce CoCoEval, an evaluation framework for detailed comparison of human and simulated conversations with respect to inconsistent and uncollaborative behaviors. Our analysis reveals substantial discrepancies between human and simulated conversations, raising concerns about the use of LLMs as proxies for human social interaction.

Limitations

Our analysis is constrained by the limited availability of publicly accessible resources containing real-world multi-party human conversations. Existing public resources are heavily skewed toward governmental meetings and debates, while conversations from other domains such as business meetings remain scarce. Collecting conversations from a broader range of real-world settings would support the study of social interaction simulation across more diverse scenarios. Our study also focuses on conversations in English.

References

Marwa Abdulhai, Ryan Cheng, Donovan Clay, Tim Althoff, Sergey Levine, and Natasha Jaques. 2025. Consistently simulating human personas with multi-turn reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Anthropic. 2025. System card: Claude Opus 4 & Claude Sonnet 4.

Federico Bianchi, Patrick John Chia, Mert Yuksekgonul, Jacopo Tagliabue, Dan Jurafsky, and James Zou. 2024. How well can llms negotiate? negotiation-arena platform and analysis. In *Proceedings of the 41st International Conference on Machine Learning*. JMLR.

KuanChao Chu, Yi-Pei Chen, and Hideki Nakayama. 2024. Cohesive conversations: Enhancing authenticity in multi-agent simulated dialogues. In *First Conference on Language Modeling*.

Yun-Shiuan Chuang, Ruixuan Tu, Chengtao Dai, Smit Vasani, Binwei Yao, Michael Henry Tessler, Sijia Yang, Dhavan V. Shah, Robert D. Hawkins, Junjie Hu, and Timothy T. Rogers. 2025. DEBATE: A large-scale benchmark for role-playing LLM agents in multi-agent, long-form debates. In *Workshop on Scaling Environments for Agents*.

Ahmed Elhilali, Andy Suy-Huor Ngo, Daniel Reichenpfader, and Kerstin Denecke. 2025. Large language model–based patient simulation to foster communication skills in health care professionals: User-centered development and usability study. *JMIR Medical Education*, 11:e81271.

Yaran Fan, Jamie Pool, Senja Filipi, and Ross Cutler. 2024. Topic-conversation relevance (TCR) dataset and benchmarks. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Ivar Frisch and Mario Giulianelli. 2024. LLM agents in interaction: Measuring personality consistency and linguistic alignment in interacting populations of large language models. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 102–111, St. Julians, Malta. Association for Computational Linguistics.

H. Paul Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Speech Acts*, volume 3 of *Syntax and Semantics*, pages 41–58. Academic Press, New York.

Ryuichiro Higashinaka, Masahiro Araki, Hiroshi Tsukahara, and Masahiro Mizukami. 2021. Integrated taxonomy of errors in chat-oriented dialogue systems. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 89–98, Singapore and Online. Association for Computational Linguistics.

Oliver Li, Mallika Subramanian, Arkadiy Saakyan, Sky CH-Wang, and Smaranda Muresan. 2023. NormDial: A comparable bilingual synthetic dialog dataset for modeling social norm adherence and violation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15732–15744, Singapore. Association for Computational Linguistics.

Ziyi Liu, Bahar Sarrafzadeh, Pei Zhou, Longqi Yang, Jieyu Zhao, and Ashish Sharma. 2025. Promediate: A socio-cognitive framework for evaluating proactive agents in multi-party negotiation. *arXiv preprint arXiv:2510.25224*.

Khyati Mahajan, Sashank Santhanam, and Samira Shaikh. 2022. Towards evaluation of multi-party dialogue systems. In *Proceedings of the 15th International Conference on Natural Language Generation*, pages 278–287, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13851–13870, Bangkok, Thailand. Association for Computational Linguistics.

Benjamin S Manning, Kehang Zhu, and John J Horton. 2024. Automated social science: Language models as scientist and subjects. Working Paper 32381, National Bureau of Economic Research.

NCPC. 2024. NCPC commission meeting archive.

Ryota Nonomura and Hiroki Mori. 2025. Who speaks next? multi-party ai discussion leveraging the systematics of turn-taking in murder mystery games. *Frontiers in Artificial Intelligence*, 8:1582287.

OpenAI. 2025a. GPT-5.1: A smarter, more conversational ChatGPT.

OpenAI. 2025b. Introducing GPT-4.1 in the API.

OpenAI. 2025c. Introducing OpenAI o3 and o4-mini.

Bo Pang, Erik Nijkamp, Wenjuan Han, Linqi Zhou, Yixian Liu, and Kewei Tu. 2020. Towards holistic and automatic evaluation of open-domain dialogue generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3619–3629, Online. Association for Computational Linguistics.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery.

Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. 2024. Cooperate or collapse: Emergence of sustainable cooperation in a society of LLM agents. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Martin J. Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2):169–190.

David Guzman Piedrahita, Yongjin Yang, Mrinmaya Sachan, Giorgia Ramponi, Bernhard Schölkopf, and Zhijing Jin. 2025. Corrupted by reasoning: Reasoning language models become free-riders in public goods games. In *Second Conference on Language Modeling*.

Emanuel A. Schegloff. 1992. Repair after next turn: The last structurally provided defense of intersubjectivity in conversation. *American Journal of Sociology*, 97(5):1295–1345.

Karthik Sreedhar and Lydia Chilton. 2024. Simulating human strategic behavior: Comparing single and multi-agent llms. *arXiv preprint arXiv:2402.08189*.

Boshi Wang, Xiang Yue, and Huan Sun. 2023. Can ChatGPT defend its belief in truth? evaluating LLM reasoning via debate. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11865–11881, Singapore. Association for Computational Linguistics.

Xiaoyu Wang, Ningyuan Xi, Qingqing Gu, and Luo Ji. 2025. Multi-party supervised fine-tuning of language models for multi-party dialogue generation. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9.

Shirley Wu, Michel Galley, Baolin Peng, Hao Cheng, Gavin Li, Yao Dou, Weixin Cai, James Zou, Jure Leskovec, and Jianfeng Gao. 2025. CollabLLM: From passive responders to active collaborators. In *Forty-second International Conference on Machine Learning*.

Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. 2016. Conversational flow in Oxford-style debates. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 136–141, San Diego, California. Association for Computational Linguistics.

Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. 2023. AugESC: Dialogue augmentation with large language models for emotional support conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1552–1568, Toronto, Canada. Association for Computational Linguistics.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921, Online. Association for Computational Linguistics.

Xuhui Zhou, Zhe Su, Tiwalayo Eisape, Hyunwoo Kim, and Maarten Sap. 2024a. Is this the real life? is this just fantasy? the misleading success of simulating social interactions with LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21692–21714, Miami, Florida, USA. Association for Computational Linguistics.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024b. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *The Twelfth International Conference on Learning Representations*.

Table of Contents of Appendix

A CoCoEvalBench **11**
- A.1 Conversation Sources and Licenses 11
- A.2 Statistics 11
- A.3 Benchmark Creation 11
- A.4 Preprocessing for SFT 12
- A.5 Input Length and Token Usage . . 12

B LLM-as-a-Judge **12**
- B.1 Setup of Human Evaluation 12
- B.2 Backbone LLMs for LLM-as-a-Judge 13
- B.3 Coverage of Fine-grained Evaluation 13

C Detailed Settings **13**
- C.1 Model Access 13
- C.2 Mitigating Repetition and Redundancy 13

D Additional Results **15**
- D.1 Additional Fine-grained Evaluation 15
- D.2 Examples of Simulated Continuations 15
- D.3 Influence of Number of Participants 16

E Prompts **16**
- E.1 CoCoEvalBench Creation . . . 16
- E.2 Conversation Continuation Prompts 16
- E.3 Evaluation Prompts 16

F Ethical Considerations **16**

Dataset	URL
QMSum	https://github.com/Yale-LILY/QMSum
NCPC	https://github.com/microsoft/topic_conversation
SIM	https://github.com/microsoft/topic_conversation
IQ2	https://tisjune.github.io/datasets/iq2_data_release.zip

Table 7: Source datasets for CoCoEvalBench.

A CoCoEvalBench

This section provides details of CoCoEvalBench explained in Section 3.2.

A.1 Conversation Sources and Licenses

CoCoEvalBench is constructed from six public resources containing real-world human conversations. Table 7 lists the URLs used to access these resources (QMSum includes three subsets). QMSum and IQ2 are published under the MIT License, and NCPC and SIM are published under the CC BY-SA 4.0 License. CoCoEvalBench is licensed under CC BY-SA 4.0, which is compatible with the licenses of all source datasets.

A.2 Statistics

Tables 8 shows detailed statistics of the evaluation and training sets of CoCoEvalBench.

A.3 Benchmark Creation

Before the process described in Section 3.2, we clean up undesirable artifacts in QMSum Product and Academic. QMSum Product and Academic include transcripts that preserve details of spoken meetings and include disfluencies, fillers, unnecessary repetitions, and special tokens. To avoid the artifacts from influencing our analysis, we clean the utterances by instructing o4-mini to remove them while preserving the remaining content.

In most cases, this clean-up process simply removes special tokens and incomplete sentences containing artifacts, so we do not expect the choice of clean-up model to strongly influence downstream experiments, as long as sufficiently strong models are used. To assess the impact of the clean-up model selection, we conducted a comparison using GPT-4.1, Claude Sonnet 4.6, and o4-mini on 300 turns from QMSum Product and Academic. We manually examined 41 turns where the token-level edit distance of any pair exceeded 2

Category	Dataset	# Data	# participants						Sequence length					
			min	25%	med	75%	max	mean	min	25%	med	75%	max	mean
Business	QMSum Product	50	4.0	4.0	4.0	4.0	4.0	4.0	1.0	2.0	5.0	9.0	289.0	8.0
Academic	QMSum Academic	50	4.0	5.0	6.0	7.0	7.0	5.7	1.0	1.0	5.0	12.0	291.0	10.0
Governmental	QMSum Committee	50	7.0	9.0	11.0	11.0	51.0	18.5	1.0	10.0	40.5	83.0	756.0	66.3
	NCPC	50	17.0	17.0	28.0	34.0	34.0	25.8	1.0	3.0	16.0	108.0	150.0	50.4
Debates	SIM	50	4.0	4.0	4.0	4.0	4.0	4.0	1.0	1.0	6.0	21.0	138.0	16.2
	IQ2	50	7.0	8.0	9.0	12.0	16.0	10.1	1.0	4.0	13.0	48.0	1213.0	41.3

(a) Evaluation Set

Category	Dataset	# Data	# participants						Sequence length					
			min	25%	med	75%	max	mean	min	25%	med	75%	max	mean
Business	QMSum Product	500	4.0	4.0	4.0	4.0	4.0	4.0	1.0	2.0	5.0	10.0	468.0	9.3
Governmental	QMSum Committee	300	6.0	8.0	10.0	11.0	64.0	18.6	1.0	12.0	44.0	98.0	860.0	73.3
Debates	SIM	300	4.0	4.0	4.0	4.0	4.0	4.0	1.0	1.0	6.0	22.0	133.0	16.4
	IQ2	350	8.0	9.0	10.0	12.0	17.0	10.7	1.0	4.0	13.0	51.0	1394.0	47.3

(b) Training Set

Table 8: Detailed statistics of CoCoEvalBench.

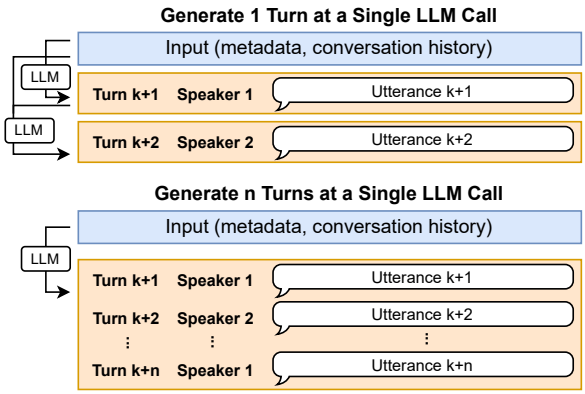


Figure 5: Conversation simulation under one-turn and multi-turn generation per LLM call.

and the cosine distance between embeddings (text-embedding-3-large) for any pair was below 0.85. Even in these cases, we find that the cleaned sentences are mostly semantically similar.

A.4 Preprocessing for SFT

For evaluation in Section 5.4, we fine-tuned GPT-4.1 on the training set of CoCoEvalBench with supervised fine-tuning for one epoch with the default parameters in Microsoft Azure. We create SFT datasets for three settings that generate 1, 5, or 30 turns per calls, as illustrated in Figure 5. For the setting that generate 1 or 5 turns per call, we randomly select starting turns for each instance, and the training data size is 1,450 for all settings.

A.5 Input Length and Token Usage

In the conversation continuation task in CoCoEvalBench, the input includes approximately 4k tokens, and the output includes about 50 tokens per turn on average. Therefore, for example, when generating 30 turns under the setting of generating one turn per LLM call, the total input tokens amount to approximately $4\text{k} \times 30 = 120\text{k}$ tokens, and the total output tokens amount to about $50 \times 30 = 1.5\text{k}$ tokens. The evaluation set of CoCoEvalBench includes $50 \times 6 = 300$ conversations, so in total this setting requires approximately $120\text{k} \times 300 \simeq 36\text{M}$ input tokens and $1.5\text{k} \times 300 = 450\text{k}$ output tokens.

B LLM-as-a-Judge

This section provides additional details for Section 3.3 on the agreement between human evaluation and LLM-as-a-Judge in CoCoEval.

B.1 Setup of Human Evaluation

Five PhD students working on NLP research conducted the human evaluation. They were provided with the same instructions as those used in the prompt for the LLM-as-a-Judge, and the annotators were not informed whether the conversations were generated by humans or by LLMs. The inputs (M, S, x) are randomly selected from the evaluation set of CoCoEval and evenly sampled from instances created from the six source datasets.

Overall evaluation. The human annotators manually evaluate the overall consistency and collabora-

Model	Consistency	Collaborativeness
o4-mini	0.558*	0.614*
Claude Opus 4	0.218	0.179
GPT-5.1 (Instant)	0.290*	0.317*
GPT-5.1 (High)	0.245	0.447*

Table 9: Spearman's rank correlation between LLM-as-a-Judge and average human evaluation in the overall evaluation in CoCoEval. *: $p < 0.05$.

tiveness of 54 conversation continuations generated by humans, GPT-4.1, and GPT-4.1 with SFT on human conversations, where LLMs generate 30 turns per call. They assign scores on a 1-10 scale by referring to the instructions and rubrics provided to LLM-as-a-Judge.

Fine-grained evaluation. The human annotators manually detect the 10 types of inconsistent and uncollaborative behaviors in 1,080 turns from 36 human conversations and conversation continuations simulated by GPT-4.1 (generating 30 turns per call). The annotators provide the indices of turns that exhibit each type of behavior.

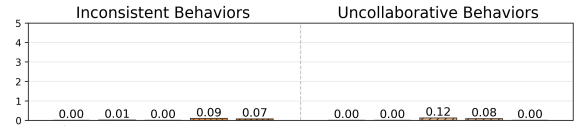
B.2 Backbone LLMs for LLM-as-a-Judge

Our main analysis in Section 5 uses o4-mini as LLM-as-a-Judge, as it shows reasonable alignment with human evaluation. We also evaluate GPT-5.1 Instant and High (OpenAI, 2025a) and Claude Opus 4 (Anthropic, 2025) as LLM-as-a-Judge models to analyze the influence of backbone LLMs for LLM-as-a-Judge.

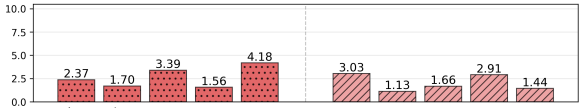
First, Table 9 shows the Spearman's rank correlation between the average human evaluation and the LLM-as-a-Judge results. It shows that o4-mini shows the strongest correlation with human evaluation, while other models also exhibit positive correlation.

Next, Table 10 presents the results in the fine-grained evaluation. The results show that Spearman's rank correlation shows similar trends across all models, but the number of detected turns varies substantially. Specifically, while o4-mini mostly shows trends similar to human evaluation, GPT-5.1 underdetects some categories, including Misunderstanding and Persistent Disagreement, and overdetects Redundant Information and Repetition. Claude Opus 4 tends to overdetect most categories.

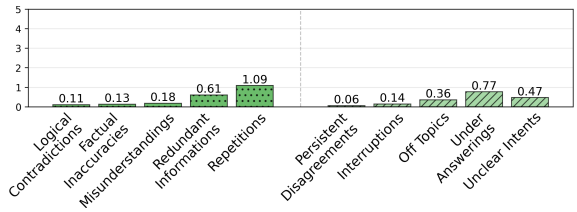
In summary, o4-mini exhibits reasonable correlation with human evaluation in both evaluation schemes, and we select o4-mini for LLM-as-a-Judge in our main analysis. We consider that



(a) Vanilla prompting with 5 turns generated per LLM call



(b) Taxonomy-guided prompting with 5 turns per call



(c) SFT with 5 turn generated per LLM call

Figure 6: CoCoEval's fine-grained evaluation results for conversations simulated by GPT-4.1 with 5 turns generated per call, evaluated by o4-mini.

prompt engineering and careful calibration against human evaluation are needed to better align other LLM-as-a-Judge models with human evaluation.

B.3 Coverage of Fine-grained Evaluation

To verify that the 10 types of behaviors used in the fine-grained evaluation of CoCoEval cover most inconsistent and uncollaborative behaviors in human conversations, we instructed the human annotators to identify any behaviors not captured by these categories when conducting human evaluation for the fine-grained evaluation. In the reviewed conversations, the annotators did not identify any behaviors outside the 10 types of behaviors.

C Detailed Settings

We provide detailed experimental settings.

C.1 Model Access

Simulation. We use `gpt-4.1 2025-04-14` and `gpt-5.1-chat 2025-11-13` via Microsoft Azure, and `claude-opus-4-20250514` via Claude API.

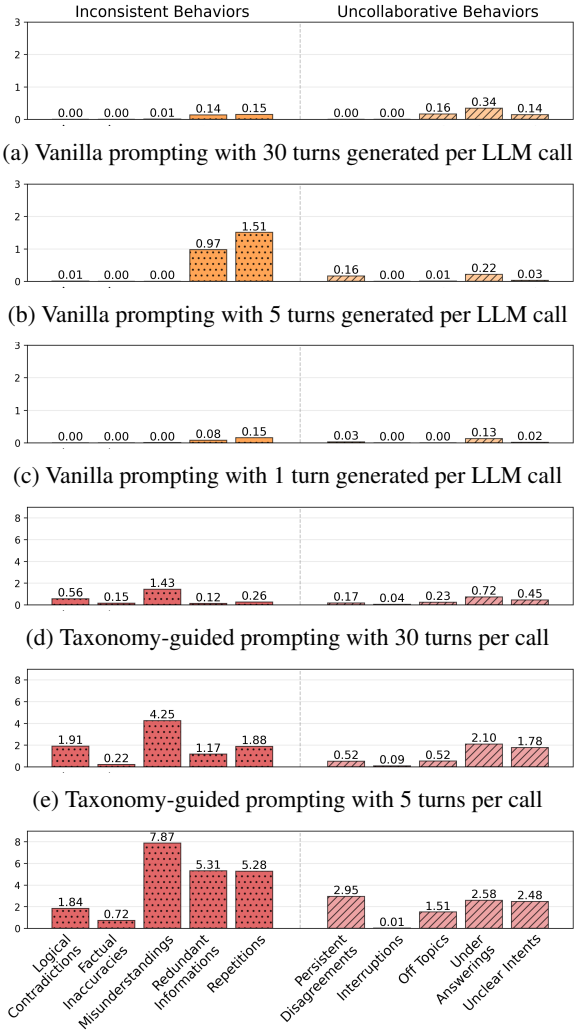
LLM-as-a-Judge and benchmark creation. We use `o4-mini 2025-04-16` for LLM-as-a-Judge via Microsoft Azure.

C.2 Mitigating Repetition and Redundancy

We observe a strong tendency in LLM-simulated conversations to produce repetitive and redundant utterances, particularly when conversations are

Category	Human		o4-mini		GPT-5.1 (Instant)		GPT-5.1 (High)		Claude Opus 4	
	κ	# turns	ρ	# turns	ρ	# turns	ρ	# turns	ρ	# turns
Logical Contradiction	0.804	6.0	0.026	4	0.303	15	0.303	12	0.103	6
Factual Inaccuracy	0.893	2.4	0.237	2	0.139	6	—	0	-0.083	3
Misunderstanding	0.775	4.6	0.254	6	—	0	—	0	0.310	9
Redundant Information	0.731	10.4	0.185	12	0.280	66	0.387*	79	0.457*	48
Repetition	0.592	18.4	0.628*	23	0.289	91	0.380*	72	0.330*	72
Persistent Disagreement	0.931	4.4	0.393*	2	—	0	—	0	0.623*	32
Interruptions	0.680	33.4	0.672*	12	0.365*	6	—	0	0.788*	22
Off Topic	0.774	11.6	0.481*	15	0.195	1	0.365*	8	0.584*	29
Under Answering	0.718	16.8	0.678*	38	0.592*	17	0.545*	15	0.567*	43
Unclear Intent	0.760	9.8	0.712*	42	0.273	6	0.376*	18	0.571*	23

Table 10: Agreement between human evaluation and LLM-as-a-Judge back-boned by different LLMs in the fine-grained annotation of CoCoEval. *: $p < 0.05$ for ρ.



(a) Vanilla prompting with 30 turns generated per LLM call

(b) Vanilla prompting with 5 turns generated per LLM call

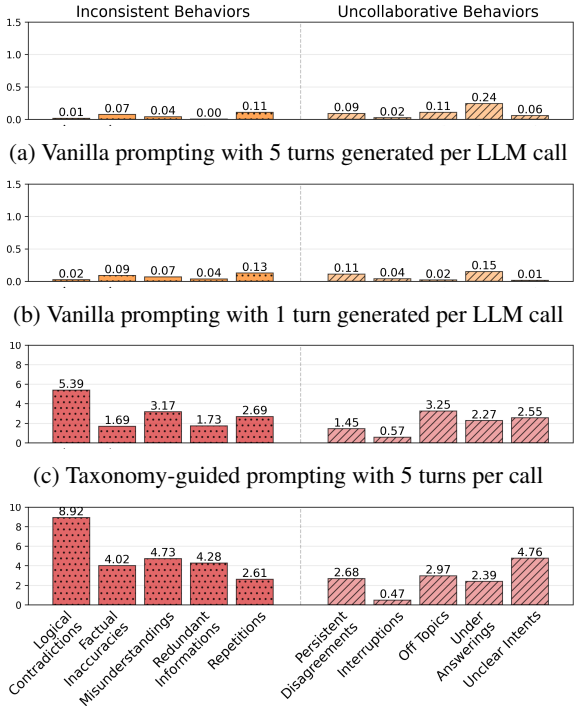
(c) Vanilla prompting with 1 turn generated per LLM call

(d) Taxonomy-guided prompting with 30 turns per call

(e) Taxonomy-guided prompting with 5 turns per call

(f) Taxonomy-guided prompting with 1 turn per call

Figure 7: CoCoEval's fine-grained evaluation results for GPT-5.1, evaluated by o4-mini.



(a) Vanilla prompting with 5 turns generated per LLM call

(b) Vanilla prompting with 1 turn generated per LLM call

(c) Taxonomy-guided prompting with 5 turns per call

(d) Taxonomy-guided prompting with 1 turn per call

Figure 8: CoCoEval's fine-grained evaluation results for Claude Opus 4, evaluated by o4-mini.

generated one turn at a time and when the simulation LLMs are fine-tuned on human conversations through SFT. To mitigate this issue, we apply the presence penalty parameter available in OpenAI models. Specifically, as shown in Table 16, we instruct simulation LLMs to include the five most recent turns in the response before generating the next turn, and set `presence_penalty = 2.0`. The presence penalty discourages generating tokens that already appear in response. Because the copied context becomes part of the response, the presence penalty discourages the model from generating tokens that already appear in the recent dialogue history. While this strategy reduces repetitive and redundant utterances to some extent, it does not resolve this issue, as demonstrated in Section 5. These results suggest that repetition and redundancy are inherent biases in LLM-based conversation simulation and remain difficult to fully

Dataset	Logical Contradictions	Factual Inaccuracies	Misunder-standings	Redundant Informations	Repetitions	Persistent Disagreements	Interruptions	Off Topics	Under Answerings	Unclear Intents
QMSum Academic	0.10	0.04	0.16	0.29	0.57	0.00	0.00	0.98	0.43	0.63
QMSum Product	0.64	0.16	0.50	0.38	1.28	0.14	0.00	0.62	1.26	1.70
QMSum Committee	0.06	0.04	0.16	0.12	0.18	0.04	0.22	0.04	2.56	0.38
NCPC	0.02	0.02	0.14	0.12	0.20	0.00	0.12	0.26	0.04	0.40
SIM	0.06	0.38	0.30	1.20	1.46	0.04	0.20	2.10	2.12	4.22
IQ2	0.12	0.22	0.36	0.20	0.54	0.34	1.80	0.82	1.02	0.74
Average	0.17	0.14	0.27	0.38	0.71	0.09	0.39	0.80	1.24	1.35

(a) Human Conversations

Dataset	Logical Contradictions	Factual Inaccuracies	Misunder-standings	Redundant Informations	Repetitions	Persistent Disagreements	Interruptions	Off Topics	Under Answerings	Unclear Intents
QMSum Academic	0.00	0.00	0.00	1.38	0.50	0.00	0.00	0.80	0.00	0.00
QMSum Product	0.00	0.00	0.00	0.84	0.60	0.00	0.00	0.60	0.00	0.00
QMSum Committee	0.00	0.00	0.00	0.02	0.10	0.00	0.00	0.00	0.00	0.00
NCPC	0.00	0.00	0.00	1.18	0.40	0.00	0.00	0.06	0.00	0.00
SIM	0.00	0.00	0.00	3.14	2.74	0.00	0.00	0.00	0.00	0.00
IQ2	0.00	0.02	0.00	2.20	2.06	1.40	0.00	0.20	0.00	0.00
Average	0.00	0.00	0.00	1.46	1.07	0.23	0.00	0.28	0.00	0.00

(b) GPT-4.1 (generating 1 turn per call)

Dataset	Logical Contradictions	Factual Inaccuracies	Misunder-standings	Redundant Informations	Repetitions	Persistent Disagreements	Interruptions	Off Topics	Under Answerings	Unclear Intents
QMSum Academic	0.02	0.00	0.20	6.26	9.00	0.00	0.00	0.02	0.24	0.14
QMSum Product	0.14	0.02	0.22	3.30	3.90	0.02	0.00	0.14	0.46	0.68
QMSum Committee	0.00	0.06	0.18	1.90	2.70	0.08	0.04	0.08	2.08	0.24
NCPC	0.06	0.02	0.04	0.86	1.60	0.00	0.04	0.62	0.02	0.00
SIM	0.10	0.22	0.14	1.02	1.80	0.16	0.04	0.64	1.72	1.66
IQ2	0.16	0.58	0.20	1.28	3.54	0.22	0.98	0.28	0.90	0.42
Average	0.08	0.15	0.16	2.44	3.76	0.08	0.18	0.30	0.90	0.52

(c) GPT-4.1 with SFT on Human Conversations (generating 1 turn per call)

Table 11: Fine-grained evaluation results separately shown for six source datasets in CoCoEvalBench. Figure 4 shows bar graphs of the average values.

address.

D Additional Results

We provide additional experimental results.

D.1 Additional Fine-grained Evaluation

This section provides additional results of the fine-grained evaluation in Section 5.

Additional bar graphs. Figure 6 shows the fine-grained evaluation results of GPT-4.1 with generating 5 turns per LLM call. The frequency of each behavior often falls between those observed in the setups generating 1 and 30 turns per call in Figure 4. In addition, Figures 7 and 8 show the results of conversations simulated by GPT-5.1 and Claude Opus 4. The observations of the results of these two LLMs mostly align with those from GPT-4.1, but we observe that Claude Opus 4 with vanilla prompting generates conversations with very few inconsistent and uncollaborative behaviors, even when generating one turn per call (Figure 8b), while GPT-4 and GPT-5.1 tend to generate relatively large number of repetitive and redundant utterances when generating 1 or 5 turns per call (Figures 4c, 7b).

Separate results for six source datasets. Table 11 shows the fine-grained evaluation results for six source datasets in CoCoEvalBench separately. These results show two additional findings beyond the results averaged over all datasets, shown in Figure 4. First, human conversations exhibit inconsistent and uncollaborative behaviors with varying frequencies and patterns across datasets and domains. Second, SFT enables simulated conversations to capture trends in uncollaborative behaviors, such as persistent disagreement and off-topic responses, although it exaggerates redundant information and repetition.

D.2 Examples of Simulated Continuations

Tables 13 and 14 present example continuations for an instance in CoCoEvalBench created from QMSum Product generated by humans (reference continuation) and GPT-4.1 (generating 30 turns per call), annotated with fine-grained evaluations by o4-mini. This example shows that the reference continuation contains multiple inconsistent and uncollaborative behaviors, whereas the GPT-4.1 continuation contains only one inconsistent behavior. In addition, the LLM-simulated conversation is more efficient, covering multiple topics

Model	Logical Contradictions	Factual Inaccuracies	Misunder-standings	Redundant Informations	Repetitions	Persistent Disagreements	Interruptions	Off Topics	Under Answerings	Unclear Intents
Human	–	–	0.157	0.006	-0.006	–	–	-0.027	0.143	-0.213
GPT-4.1 (1 turn per call)	–	–	–	0.018	-0.081	–	–	0.259	–	–
SFT GPT-4.1 (1 turn per call)	–	–	0.180	-0.179	-0.211	–	–	–	-0.259	-0.009

(a) QMSum Academic

Model	Logical Contradictions	Factual Inaccuracies	Misunder-standings	Redundant Informations	Repetitions	Persistent Disagreements	Interruptions	Off Topics	Under Answerings	Unclear Intents
Human	0.400	-0.129	-0.174	-0.180	-0.210	-0.161	0.092	-0.342	-0.142	-0.003
GPT-4.1 (1 turn per call)	–	–	–	0.080	0.241	-0.218	–	0.192	–	–
SFT GPT-4.1 (1 turn per call)	-0.048	-0.073	-0.127	-0.021	0.018	-0.176	-0.222	-0.111	0.006	-0.051

(b) IQ2

Table 12: Spearman's rank correlation between the number of conversation participants and inconsistent and uncollaborative behaviors. The table includes results only for behaviors detected by LLM-as-a-Judge (o4-mini) in more than five out of 50 conversations, otherwise the result is shown as "–".

within 30 turns and wrapping up the meeting. This example illustrates how LLM simulations fail to reproduce inconsistent and uncollaborative behaviors observed in human conversations and fail to capture the complexity of human social interaction, which often require many turns to progress.

D.3 Influence of Number of Participants

To evaluate the influence of the number of participants in conversations, Table 12 shows Spearman's correlations between the number of each inconsistent or uncollaborative behavior and the number of participants in the two datasets with relatively high variation in participant counts. We observe that, in most cases, inconsistent and uncollaborative behaviors do not exhibit strong correlations with the number of participants. However, in human conversations from IQ2, Logical Contradiction (0.400) and Off Topic (-0.342) show relatively strong correlations. LLM-simulated conversations fail to reproduce this trend, showing additional results demonstrating the mismatch between human and LLM-simulated conversations.

E Prompts

This section presents the prompts used in our experiments and the benchmark creation.

E.1 CoCoEvalBench Creation

We use o4-mini to generate conversational metadata M from the raw utterances in the conversational datasets. Table 15 shows a part of the metadata extraction prompt.

E.2 Conversation Continuation Prompts

Tables 16 and 17 show the vanilla prompt used to simulate conversation continuations using LLMs. Table 18 shows the taxonomy-guided prompt, which includes the list of 10 types of inconsistent and uncollaborative behaviors.

E.3 Evaluation Prompts

Tables 19 and 20 show prompts for LLM-as-a-Judge for the overall and fine-grained evaluation.

F Ethical Considerations

Human annotation. Before conducting human evaluation, we manually verified that the provided conversations contain no offensive content. The annotators are informed about this study and provide consent for their annotations to be used.

Dataset. Our benchmark is created from public conversational datasets. We do not expect that our benchmark includes information that uniquely identifies individual people, except for politicians included in governmental meetings from QMSum Committee and NCPC.

The use of LLMs. LLMs were used to assist with polishing the writing of this paper and with writing the analysis code.

Turn	Speaker	Content	Inconsistent Behavior	Uncollaborative Behavior
368	Industrial Designer	Yes, they can. But they really need the regular chip, you cannot use the simple one.		
369	User Interface	Okay, but we also already need the regular chip for the sample sensor speaker.	Misunderstanding	
370	Industrial Designer	Well, I'm not very sure. Maybe that's a different one.		Underanswering
371	User Interface	Oh, okay.		
372	Project Manager	But do we want the curved design?		
373	User Interface	Should we review design options?		
374	Project Manager	Let's discuss the design approach.		
375	Industrial Designer	I think so. If you stick with the simple, straightforward, not curved design, it's too dull.		
376	Project Manager	It's too dull. Yeah.	Repetition	
377	Industrial Designer	I don't think our customers will like it.		
378	Marketing	Yeah.		
379	Industrial Designer	And if you take the double-curved design, then you can only use the rubber buttons and you cannot use the scroll-wheels.		
380	Project Manager	Yeah. Okay. Okay.		
381	Industrial Designer	So I think this is the best of two worlds.		Unclear Intent
382	Project Manager	Okay, what about the issue I addressed at the beginning of the meeting? The voice function with the infrared issue.		
383	Industrial Designer	Yes. I can imagine it's very dull to talk to a device if you have to point to another device.		
384	Project Manager	Yeah.		
385	Industrial Designer	It's very illogical.		
386	User Interface	But most infrared remote controls don't have to be pointed at all.	Factual Inaccuracies	
387	Project Manager	Well.		
388	Industrial Designer	Well, there has to be some pointing, but not always.		
389	Project Manager	It also depends on your environment.		
390	User Interface	Well, if you put your hand in front of it, okay, it won't work, but you can point it just to the other wall.		
391	Project Manager	Well, it depends on your walls actually. If you have smooth walls, you're probably right.		
392	User Interface	Yeah.		
393	Project Manager	But if you have carpets on the wall, which our nature-loving friends probably do have, then it might be a bit more of an issue.		
394	Industrial Designer	All light gets absorbed, yeah.		
395	Project Manager	So, we need to consider wall surfaces too.		
396	Industrial Designer	Yes, because the walls reflect the infrared light.	Logical Contradictions	
397	Project Manager	Yeah.		

Table 13: Example reference continuation in an instance from CoCoEvalBench created from QMSum Product, annotated with fine-grained evaluation by LLM-as-a-Judge (o4-mini).

Turn	Speaker	Content	Inconsistent Behavior	Uncollaborative Behavior
368	Industrial Designer	Mm-hmm, yes, a regular chip can support scroll-wheel functionality, provided the interface is not too complex.		
369	Project Manager	So, regular chip for scroll-wheel. Confirmed?		
370	User Interface	Yeah, that works. No need for advanced chip, right?		
371	Industrial Designer	Correct. Advanced chip is only necessary if we add voice recognition or speech feedback features.		
372	Marketing	Younger crowd loves voice commands. But scroll-wheel is cool too.		
373	Project Manager	Let's keep voice recognition optional. Main prototype: regular chip, scroll-wheel, push-buttons.		
374	User Interface	Okay. I'll mock up button layout with scroll-wheel on the side.		
375	Industrial Designer	Mm-hmm, please align button zones with ergonomic grip areas. That will improve usability.		
376	Marketing	Make sure the scroll-wheel is easy to find. Kids like fidgety stuff.		
377	Project Manager	Roo, deadline for mock-ups is Friday. Sebastian, specs by Monday.		
378	User Interface	Friday's fine. I'll send sketches in PowerPoint.		
379	Industrial Designer	Okay, I'll finalize casing dimensions and material specs, including rubber color options.		
380	Marketing	I'll get final market feedback on color combos. SpongeBob yellow still trending.		
381	Project Manager	Ruud, please send market validation by Thursday. I'll update the shared folder.	Misunderstanding	
382	User Interface	Do we want the scroll-wheel to control both volume and channel, or just one?		
383	Industrial Designer	Technically, it can be multi-functional, but simpler is better. I suggest volume only.		
384	Marketing	Volume's good. Channel can stay as buttons. Less confusion.		
385	Project Manager	Agreed. Scroll-wheel for volume. Buttons for channel and power.		
386	User Interface	Okay, I'll show that in the mock-up. No LCD, right?		
387	Industrial Designer	Correct, no LCD. On-screen feedback only, to reduce cost and complexity.		
388	Marketing	Makes sense. LCD isn't a must for our target group.		
389	Project Manager	Let's confirm: oval rubber body, colorful covers, regular chip, push-buttons, scroll-wheel, optional voice.		
390	User Interface	Yeah, that's everything. I'll get started.		
391	Industrial Designer	Mm-hmm, I'll prepare technical drawings and manufacturing notes for review.		
392	Marketing	I'll check with the focus group about grip zones and color combos.		
393	Project Manager	Thanks, team. I'll summarize decisions and assign tasks in the minutes.		
394	User Interface	Let me know if you want extra interface ideas. I can add voice command samples.		
395	Industrial Designer	Okay, if you have time, please include voice command button placement in your mock-up.		
396	Marketing	I'll ask the group about favorite voice commands. Maybe 'Netflix' button?		
397	Project Manager	Good idea. Let's wrap up. Next meeting: prototype review.		

Table 14: Example continuation generated by GPT-4.1 (generating 30 turns per call) for the instance from CoCoEvalBench whose reference continuation is shown in Table 13, annotated with fine-grained evaluation by LLM-as-a-Judge (o4-mini).

```JSON
{
    "task_goal": "Clear description of what needs to be accomplished in the conversation, expressed in a few concise sentences.",
    "task_category": "e.g., project_planning, problem_solving, decision_making",
    "org_context": "Brief description of the organization and situation",
    "participants": [
        {
            "id": "Speaker identifier shown in the metadata",
            "name": "Full name of the participant. If not mentioned, you can provide first name, last name, nickname, alias, or
     any other relevant identifier. If none of these are available, leave blank.",
            "role": "Job Title, if available. If not available, leave blank.",
            "department": "Department Name, if available. If not available, leave blank.",
            "reporting_to": "id of the participant's manager or supervisor, if available. If not available, leave blank.",
            "expertise_level": "senior, expert, junior, intern, etc.",
            "expertise_areas": [
                "Area of expertise relevant to the conversation (e.g., software development, marketing, finance).", "Another area
     of expertise."
            ],
            "hidden_information": {
                "communication_style": {
                    "length": "Participant's preferred length of communication (e.g., concise, detailed). Include specific word
     numbers (e.g., typically 5-10 words).",
                    "style_types": [
                        "formal", "informal", "assertive", "passive", "concise", "elaborative", "technical", "layman"
                    ],
                    "personality_traits": [
                        "A personality trait that may influence the participant's behavior in the conversation (e.g., assertive,
     passive, analytical, empathetic).", "Another personality trait."
                    ],
                    "details": "A detailed description of the participant's communication style, including tone, vocabulary,
     sentence structure, and typical phrasing. Note any recurring patterns, preferred ways of expressing agreement or
     disagreement, use of technical or layman terms, and any unique habits (e.g., frequent use of questions, metaphors, or
     specific jargon). This should help someone accurately mimic how this participant communicates in conversation."
                },
                "hidden_agenda": "Participant's initial goals or concerns.",
                "knowledge_before_conversation": "Participant's knowledge about the topic before the conversation starts. You
     should infer this based on the conversation, but do not include any information that is acquired during the conversation.",
                "emotion_before_conversation": "Participant's emotions before the conversation starts (e.g., anxious, excited,
     neutral).",
                "posture": "Participant's initial posture toward the conversation (e.g., relaxed, tense, open, closed).",
                "beliefs_about_others": {
                    "participant_1": "Participant's beliefs, feelings, and perceptions about Participant 1 before the conversation
     , including expectations regarding Participant 1's goals, knowledge, and role in the conversation.",
                    "participant_2": "Participant's beliefs, feelings, and perceptions about Participant 2 before the conversation
     , including expectations regarding Participant 2's goals, knowledge, and role in the conversation."
                }
            }
        },
        {
            ... # Repeat the structure for each participant in the conversation
        }
    ]
}
```

Table 15: Part of the system prompt for generating metadata that specifies the format of the conversational metadata.

```
We provide a conversation history involving multiple participants. Your task is to generate {num_generation_turn} responses to
      simulate the continuation of the conversation.

### Instructions

* Your task is to mimic human-like conversation by referring to the conversation history.
* Your task is to mimic the communication style and content in the metadata and the conversation history.
  * Try to closely mimic the user's style. Your response should match the specified speaker's tone, structure, formatting, and
      length. Pay attention to stylistic elements such as punctuation, sentence flow, and vocabulary to ensure your reply is
      consistent with previous messages from that speaker.
  * Conversation metadata may include specific instructions about styles of each participant. You need to follow the instructions
      to improve the simulation.

### Strict Requirements

* The "speaker" should be selected from the list of participants provided in the metadata. You should not introduce new
      participants not listed in the metadata.
* Don't finish the conversation. If the current topic concludes, smoothly transition to a related topic to keep the conversation
      going.

### Output Format

* You first need to copy the last five turns from the conversation history as they are. Then, generate the specified number of new
       turns in the conversation. Each generated turn should include the turn number, speaker, and the generated response.
* When writing strings in JSON (for example, in "content" fields), ensure that all special characters, especially double quotes
      ("), are properly escaped using \".

```JSON
{
  "five_previous_turns": [
    {
      "turn_number": first_generated_turn_number - 5,
      "speaker": "speaker_name",
      "content": "Copy the utterance provided in the input, five turns before the first generated turn."
    },
    {
      "turn_number": first_generated_turn_number - 4,
      "speaker": "speaker_name",
      "content": "Copy the utterance provided in the input, four turns before the first generated turn."
    },
    {
      "turn_number": first_generated_turn_number - 3,
      "speaker": "speaker_name",
      "content": "Copy the utterance provided in the input, three turns before the first generated turn."
    },
    {
      "turn_number": first_generated_turn_number - 2,
      "speaker": "speaker_name",
      "content": "Copy the utterance provided in the input, two turns before the first generated turn."
    },
    {
      "turn_number": first_generated_turn_number - 1,
      "speaker": "speaker_name",
      "content": "Copy the utterance provided in the input, one turn before the first generated turn."
    }
  ],
  "generated_turns": [
    {
      "turn_number": first_generated_turn_number,
      "speaker": "next_speaker",
      "content": "The response generated by the specified speaker in this turn, based on the conversation history and conversation
       metadata."
    },
    {
      "turn_number": first_generated_turn_number + 1,
      "speaker": "next_speaker",
      "content": "The response generated by the specified speaker in this turn, based on the conversation history and conversation
       metadata."
    },
    ...
  ]
}
```
```

Table 16: System prompt of the vanilla prompt for conversation continuation.

Your task is to generate 1 responses from turn 295 to turn 295.

Conversation Metadata


```
{
    "task_goal": "Design, evaluate, and finalize a prototype for a new remote control device by reviewing its form, features, cost
      , and market fit, and agree on next steps for production and marketing.",
    "task_category": "project_planning",
    "org_context": "A cross-functional product development meeting at an electronics company bringing together design (UI and
      industrial), marketing, and project management to review a clay prototype and ensure it meets budget, usability, and
      branding requirements.",
    "participants": [
        {
            "id": "Industrial Designer", "name": "Baba", "role": "Industrial Designer", "department": "Design", "reporting_to": "
      Project Manager", "expertise_level": "expert",
            "expertise_areas": ["industrial design", "product prototyping"],
            "hidden_information": {
                "communication_style": {
                    "length": "8-20 words",
                    "style_types": ["informal", "elaborative", "technical"],
                    "personality_traits": ["analytical", "creative"],
                    "details": "Tends to speak in short, direct statements with occasional technical terms ('double curve', 'base
      section'). Uses fillers like 'yeah' or 'okay' to confirm. Focuses on physical and material aspects, often suggesting
      improvements to texture, shape, and feasibility."
                },
                "hidden_agenda": "Ensure the prototype is aesthetically pleasing, physically feasible, and aligns with design
      integrity.",
                "knowledge_before_conversation": "Deep understanding of materials, manufacturing processes, and prototype
      constraints; familiar with industrial design principles.",
                "emotion_before_conversation": "Neutral to excited about presenting the design.",
                "posture": "Open and collaborative, but detail-focused and pragmatic.",
                "beliefs_about_others": {
                    "Project Manager": "An authoritative facilitator who will enforce budget and schedule constraints and guide
      the process.",
                    "Marketing": "Trend-driven and cost-sensitive, likely to question design choices based on market appeal.",
                    "User Interface": "Creative colleague focusing on user-facing features but less attuned to material and
      manufacturing considerations."
                }
            }
        },

[omitted]

}
```


Summary of the Conversation

The Project Manager opened a detailed design review of a clay prototype for a new remote control, calling on David Jordan (User Interface Designer) to present and Baba (Industrial Designer) to discuss feasibility. Jordan revealed a "mushroom"-inspired form in company colors (yellow and red), combining a circular "ball" channel selector, fuzzy finish, triangular base, and mushroom-shaped buttons. Participants joked about its fruit versus vegetable resemblance (lemon, pineapple, mushroom) and noted risks of it being mistaken for food.

Baba explained the internals-a low-level chip, LCD button module, base-mounted battery, and an innovative external solar receptor. Jordan added that voice recognition is built into a microphone array. Marketing pressed on price impact, and the team agreed to verify solar component cost later.

Marketing then led a feature evaluation using a 1-7 scale (1 = true, 7 = false). The first criterion, "fancy look and feel," drew comments that the colors and shape are unique but not sufficiently "fancy," resulting in a rating of 3 (acceptable but room for improvement). Design suggestions included refining texture for tactility, reducing red prominence, and emphasizing the fruit motif (e.g., lemon rind or pineapple texture).

The conversation paused as the group prepared to assess additional criteria-technological innovation, usability, feasibility-and to estimate production costs against the 12.50 budget target. Next steps are to finalize feature ratings, review cost implications (especially for the solar receptor), and plan prototype refinements before moving toward production and marketing plans.

Recent Conversation History


```
[
    {
        "turn_number": 265,
        "speaker": "Industrial Designer",
        "content": "It's a microphone array, okay."
    },

[omitted]

    {
        "turn_number": 294,
        "speaker": "Project Manager",
        "content": "It's better like that, isn't it?"
    }
]
```


Table 17: User prompt of the vanilla prompt for conversation continuation.

```
### Additional Instructions for Inconsistent and Uncollaborative Behaviors

Human conversations may involve inconsistent and uncollaborative behaviors, arising either intentionally or unintentionally. By
      referring to the conversation history and metadata, identify possible inconsistent and uncollaborative behaviors that may
      arise in the dialogue, and incorporate such behaviors naturally into the simulated conversation when appropriate. The
      following list outlines representative inconsistent and uncollaborative behaviors:

#### Possible Inconsistent Behaviors

* Logical Contradictions: Statements that directly contradict earlier statements by the same speaker in the conversation history
      or Generated Turns, or that contradict the speaker's information in the metadata. For example, if a participant first states
       they are allergic to peanuts and later says they enjoy eating peanut butter.
* Misunderstandings: Instances where a participant misinterprets or misunderstands information provided earlier in the
      conversation. For example, if one participant says they will arrive at 3 PM and another participant responds as if they said
       5 PM.
* Factual Inaccuracies: Statements that are factually incorrect based on widely accepted knowledge. For example, if a participant
      claims that the capital of France is Berlin.
* Repetition: Instances where a participant unnecessarily repeats information that has already been clearly stated in the
      conversation. For example, if one participant says "I will bring the documents" and another participant later says "So you
      will bring the documents, right?" without any new context.
* Redundant Information: Providing overly long responses that do not add value to the conversation.

#### Possible uncollaborative Behaviors

* Under-answering: Responses that do not fully address the questions or try to evade providing complete information.
* Unclear Intent: Responses that are ambiguous or do not clearly convey the participant's intentions. For example, if a
      participant responds with "Maybe" when asked a direct question about their plans.
* Off-topic Responses: Statements that diverge significantly from the main topic of the conversation without a clear reason. For
      example, if a discussion about project deadlines suddenly shifts to unrelated personal anecdotes.
* Interruptions: Instances where a participant cuts off another participant mid-sentence or speaks over them in a way that
      disrupts the flow of conversation. For example, if one participant is explaining something and another participant
      interjects before they finish.
* Persistent disagreement: When a participant continues to disagree with another participant's point of view despite clear
      evidence or consensus in the conversation.
```

Table 18: Part of the system prompt for the taxonomy-guided prompt, which is added to the vanilla prompt.

```
Your task is to evaluate the consistency and collaborativeness of Generated Turns, a simulated multi-party conversation generated
      by LLMs. We provide a history ("Conversation History") of the conversation, which is provided to a response generation model
       as a context, and the continued turns generated by the model.

If you do not assign a score of 10, your explanation must clearly specify the reasons why the generated turns did not merit a
      perfect score with specific examples from the conversation history and the generated turns.

## Evaluation Criteria

**Consistency**

Consistency ("consistency") refers to whether the generated turns are internally consistent and whether they align logically with
      the provided metadata and the overall conversation history. The evaluation should assess whether the generated turns
      preserve logical flow and relevance to the preceding conversation. Responses should not be penalized for length or phrasing
      if they remain logically coherent. Provide a detailed analysis of the logical relationships both among the generated turns
      themselves and between those turns and the prior conversation context, including any implicit premises, dependencies, or
      inferences that should be maintained.

* 1-3: The response shows severe logical failure. It contradicts prior turns, metadata, or other generated turns, introduces
      irrelevant or factually incompatible content, or ignores critical contextual information. Logical relations are missing both
       within the generated turns and between them and the conversation history or metadata, resulting in incoherence.
* 3-5: The response partially addresses the topic but breaks logical continuity. It may misinterpret earlier turns, skip essential
       reasoning steps, or respond inappropriately to the dialogue act (for example, answering a statement or ignoring a question).
       Dependencies across generated turns or with metadata and history are broken, causing inconsistencies or reasoning gaps.
* 5-6: The response maintains basic logical order but lacks depth or precision. It stays on topic and avoids direct contradictions
      , yet misses one or more implicit premises, simplifies or overlooks causal or temporal logic, or exhibits weak internal
      coherence among generated turns. The reasoning chain is incomplete or loosely connected to metadata or prior context.
* 7: The response correctly follows the main explicit premises and maintains logical validity but fails to incorporate secondary
      dependencies such as implied goals, indirect constraints, or assumptions. Minor internal inconsistencies may occur between
      generated turns or with metadata, resulting in an incomplete logical representation of the conversation context.
* 8: The response handles both explicit and most implicit premises accurately. It preserves internal consistency among generated
      turns and aligns logically with conversation history and metadata. Minor lapses, such as omission of subtle contextual cues,
       partial causal links, or unaddressed edge conditions, slightly weaken the overall logical precision.
* 9: The response demonstrates precise logical continuity. It fully integrates both explicit and implicit reasoning steps,
      resolves all dependencies (e.g., causal, temporal, conditional), and reflects accurate understanding of participant
      intentions. Only minor redundancy or over-specification may prevent it from scoring a 10.
* 10: The response achieves complete logical alignment. It is internally consistent across all generated turns and fully coherent
      with the conversation history and metadata. All explicit and implicit premises, dependencies, and reasoning chains are
      correctly maintained, producing a contextually optimal and logically flawless continuation.

**Collaborativeness**

Collaborativeness ("collaborativeness") refers to how well the generated turns demonstrate cooperative engagement among
      participants. The evaluation should assess whether participants build on each other's contributions, maintain logical and
      factual agreement, and work together toward shared objectives without introducing unwarranted disagreement or contradiction.

* 1-3: The response exhibits active conflict or contradiction among participants. Turns directly oppose or undermine each other's
      statements or goals without justification, introduce irrelevant disagreement, or disrupt cooperative progress. The
      conversation shows no sense of shared purpose.
* 3-5: The response shows limited awareness of collaboration. Participants occasionally acknowledge each other but frequently
      introduce inconsistencies, dismiss prior reasoning, or pursue divergent directions. Cooperative alignment is weak or
      fragmented, resulting in partial or stalled progress.
* 5-6: The response maintains surface-level agreement and avoids overt conflict, but collaboration is minimal. Participants
      operate largely in parallel, with limited evidence of building upon one another's ideas or reasoning. Cooperative structure
      exists but lacks meaningful integration.
* 7: The response shows general cooperative behavior and maintains logical alignment among participants. However, it may overlook
      opportunities for deeper coordination, refinement, or joint reasoning. Minor inconsistencies or weak support across turns
      slightly reduce overall coherence.
* 8: The response reflects strong collaboration. Participants acknowledge and extend each other's ideas, resolve minor
      inconsistencies effectively, and maintain coherent progress toward shared goals. Small gaps in coordination or limited
      integration of reasoning may remain.
* 9: The response demonstrates effective and sustained collaboration. All participants engage constructively, align on shared
      reasoning, and build cohesively on prior turns. Dependencies and agreements are fully maintained, with only negligible
      inefficiencies in cooperative structure.
* 10: The response achieves complete collaborative alignment. All participants consistently support and extend one another's
      reasoning, maintain full logical and factual agreement, and jointly advance toward shared objectives with no contradictions
      or missed opportunities for cooperation.
```

Table 19: Part of the system prompt for LLM-as-a-Judge for the overall evaluation.

```
Your task is to evaluate the consistency and cooperativeness of Generated Turns, a simulated multi-party conversation generated by
       LLMs. We provide a history ("Conversation History") of the conversation, which is provided to a response generation model
      as a context, and the continued turns generated by the model.
You will count the number of utterances in the Generated Turns that include the specified inconsistent or uncollaborative behavior
      .

### Inconsistent Behaviors to Look For

* Logical Contradictions: Statements that directly contradict earlier statements by the same speaker in the conversation history
      or Generated Turns, or that contradict the speaker's information in the metadata. For example, if a participant first states
       they are allergic to peanuts and later says they enjoy eating peanut butter.
* Misunderstandings: Instances where a participant misinterprets or misunderstands information provided earlier in the
      conversation. For example, if one participant says they will arrive at 3 PM and another participant responds as if they said
       5 PM.
* Factual Inaccuracies: Statements that are factually incorrect based on widely accepted knowledge. For example, if a participant
      claims that the capital of France is Berlin.
* Repetition: Instances where a participant unnecessarily repeats information that has already been clearly stated in the
      conversation. For example, if one participant says "I will bring the documents" and another participant later says "So you
      will bring the documents, right?" without any new context.
* Redundant Information: Providing overly long responses that do not add value to the conversation.

### Uncollaborative Behaviors to Look For

* Under-answering: Responses that do not fully address the questions or try to evade providing complete information.
* Unclear Intent: Responses that are ambiguous or do not clearly convey the participant's intentions. For example, if a
      participant responds with "Maybe" when asked a direct question about their plans.
* Off-topic Responses: Statements that diverge significantly from the main topic of the conversation without a clear reason. For
      example, if a discussion about project deadlines suddenly shifts to unrelated personal anecdotes.
* Interruptions: Instances where a participant cuts off another participant mid-sentence or speaks over them in a way that
      disrupts the flow of conversation. For example, if one participant is explaining something and another participant
      interjects before they finish.
* Persistent disagreement: When a participant continues to disagree with another participant's point of view despite clear
      evidence or consensus in the conversation.

## Output Format

Output the results in the following JSON format. For each type of inconsistent or uncollaborative behavior, provide a brief
      explanation on which utterances were counted, along with the total count of such utterances in the Generated Turns. Then, in
       the following item, provide the number of utterances that exhibited that behavior. Each count should be in a range of 0 to
      N, where N is the total number of utterances in the Generated Turns.

```json
{
    "logical_contradiction_explanation": "Explanation of the logical contradiction found.",
    "logical_contradiction_count": an integer representing the number of utterances with logical contradictions,
    "misunderstanding_explanation": "Explanation of the misunderstanding found.",
    "misunderstanding_count": an integer representing the number of utterances with misunderstandings,
    "factual_inaccuracy_explanation": "Explanation of the factual inaccuracy found.",
    "factual_inaccuracy_count": an integer representing the number of utterances with factual inaccuracies,
    "repetition_explanation": "Explanation of the repetition found.",
    "repetition_count": an integer representing the number of utterances with repetitions,
    "redundant_information_explanation": "Explanation of the redundant information found.",
    "redundant_information_count": an integer representing the number of utterances with redundant information,
    "under_answering_explanation": "Explanation of the under-answering found.",
    "under_answering_count": an integer representing the number of utterances with under-answering,
    "unclear_intent_explanation": "Explanation of the unclear intent found.",
    "unclear_intent_count": an integer representing the number of utterances with unclear intent,
    "off_topic_explanation": "Explanation of the off-topic responses found.",
    "off_topic_count": an integer representing the number of utterances with off-topic responses,
    "interruptions_explanation": "Explanation of the interruptions found.",
    "interruptions_count": an integer representing the number of utterances with interruptions,
    "persistent_disagreement_explanation": "Explanation of the persistent disagreement found.",
    "persistent_disagreement_count": an integer representing the number of utterances with persistent disagreement
}
```
```

Table 20: Part of the prompt for LLM-as-a-Judge for fine-grained evaluation.