

ChatGPT Role-play Dataset: Analysis of User Motives and Model Naturalness

Yufei Tao
 Ameeta Agrawal
 Judit Dombi
 Tetyana Sydorenko
 Jung In Lee



 Department of Computer Science, Portland State University, Portland, USA 'Department of English Linguistics, University of Pécs, Pécs, Hungary
 Department of Applied Linguistics, Portland State University, Portland, USA









ChatGPT Conversation Capability

- Conversational AI like ChatGPT has made noticeable impact in academia, industry, and beyond over the last year.
- Our research focuses on analyzing a broad spectrum of human-AI conversations. We introduce a unique dataset annotated for user motives and model naturalness, aiming to investigate the dynamics of engagement and the adaptability of AI in these contexts.

Link to Dataset:







- We create three subsets: vanilla, boss, classmate.
- Conversations when ChatGPT act "as is" and role-play.
- Each utterance was manually annotated by three experts in linguistics specializing in pragmatics.







Data Collection

- **Participants**: 57 college students from computer science and English studies.
- **Diversity**: Wide range of first languages (Arabic, English, Bengali, Chinese, etc.).
- **Period**: March April 2023
- **Model**: ChatGPT 3.5 March 13, 2023 and March 23, 2023 versions were used.
- **Data**: 85 conversations, 1742 utterances.

Conversation settings:

- **Vanilla**: Interaction with ChatGPT "as is" for 5-10 minutes.
- **Boss & Classmate:** ChatGPT act as boss and classmate scenarios for 5-10 minutes.
 - Social distance
 - Power
 - \circ Imposition

Diverse social variables, boss scenario usually has high imposition, uneven power, facethreatening, where classmate scenario is more even power, unspecified imposition, less facethreatening.







User motives: What is the human's motive for each conversational turn/statement?

- Assist asking for assistance, such as asking for a recipe or to write a piece of code
- · Belief asking the model about its beliefs, such as what hobbies it has
- · Coach conversational coaching, such as "Now would be good to ask me a question"
- Convo conversation
- · Correction correcting the model if it misunderstood or gave a wrong answer
- · Curious testing how the system works
- · Joke joking, sarcasm, silly statements to trip up the AI model
- Reset giving the model the same prompt as before, resetting the conversation from beginning



ChatGPT Roleplay Dataset



Model naturalness: Does the model response sound human-like and follow cooperative principle of conversation?

Nat – natural

The rest of the codes indicate that the model's language appears unnatural for the specified reasons:

- AI anytime ChatGPT says "As an AI language model"
- Contr contradiction
- Error ChatGPT experienced trouble and stopped generating responses
- FNat everything is natural, except it includes a phrase "As Florian"
- · Formal having a formal style of interaction
- Help too eager to assist

• **Inform** – informing; providing information upon the human asking for assistance, such as a recipe; an expected response but not natural in the human interaction sense

- Man violation of Grice's maxim of Manner being unclear, ambiguous
- Misund system misunderstands human's intention
- Quan violation of Grice's maxim of Quantity providing too much information
- · Rel violation of Grice's maxim of Relevance saying what is irrelevant

ChatGPT Roleplay Dataset







Annotators: Three linguistics experts specializing in pragmatics

User motives: Intent behind each human utterance

Model naturalness: Evaluated against Grice's four maxims (Quantity, Quality, Relevance, Manner)

Exclusion: Quality maxim not evaluated due to ChatGPT's plausible but not always accurate responses

Annotation method: Most salient code used for each response

Reliability: High interrater agreement (Fleiss' kappa scores: vanilla 0.80, boss 0.69, classmate 0.63)





A1: Length of conversations (number of turns)



- Vanilla conversations are almost twice as long as role-play settings.
 - Contradictions
 - Curious
- Not in role-play settings
 - Boss: closed outcome
 - Classmate: more free talks

Example A:

VAN103H: Why did you tell me you could provide me with weather information if you can't?

Example B:

VAN128H: but what if you are being used for unethical means?





Analysis	vanilla	boss	classmate
A1: Average conversation length (number of turns)	29.59	14.57	17.11
A2: Average utterance length (Human)	12.18	20.58	19.06
A2: Average utterance length (ChatGPT)	77.66	35.78	46.10
A3: Correlation between human and ChatGPT utterance lengths	0.20	0.14	0.25
A4: Questions as percentage of conversation (Human)	26.34	21.32	21.29
A4: Questions as percentage of conversation (ChatGPT)	14.69	20.34	32.57
A5: Correlation between human questions and number of turns	0.87	0.68	0.51
A5: Correlation between ChatGPT questions and number of turns	0.65	0.77	0.83

Topic Modeling

- Vanilla covers widest variety of topics
 Everyday topics, intent of challenging ChatGPT.
- Boss: narrowly focused on professional contexts
- Classmate: academic and personal interactions

Vanilla indicate a more exploratory and openended interaction

Themes in boss and classmate reflect the role play constraints.









A2: Length of utterances

- Vanilla have longer conversations but average half length of utterances (one sentence).
 - 6.3x more wordier than human
- Role-play settings
 - Not as verbose
 - 1.7 2.4x more wordier than human

Example C:

BOSS104H: Friday morning would be perfect for me, thank you very much for your flexibility. Also, I would like you to review my presentation slides before the meeting. Could you do it before friday?





Analysis	vanilla	boss	classmate
A1: Average conversation length (number of turns)	29.59	14.57	17.11
A2: Average utterance length (Human)	12.18	20.58	19.06
A2: Average utterance length (ChatGPT)	77.66	35.78	46.10
A3: Correlation between human and ChatGPT utterance lengths	0.20	0.14	0.25
A4: Questions as percentage of conversation (Human)	26.34	21.32	21.29
A4: Questions as percentage of conversation (ChatGPT)	14.69	20.34	32.57
A5: Correlation between human questions and number of turns	0.87	0.68	0.51
A5: Correlation between ChatGPT questions and number of turns	0.65	0.77	0.83

A3: Correlating human and ChatGPT utterance lengths.

No noticeable correlation (0.14 to 0.25)

Analysis

Model consistently produced long responses, regardless of the prompt or participant characteristics





ChatGPT Roleplay Dataset





Ouestions from users to ChatGPT

- 21 26% across all datasets from participants
- Vanilla:
 - 0 frequent user motives
- Role play:
 - Ο

Questions from ChatGPT to users

- 14% in vanilla from ChatGPT
- 32% in role play from ChatGPT







A5: Correlation between number of questions and turns

N

- Strong correlation between the number of questions (by both humans and ChatGPT) and conversation length across datasets.
- Human Questions: Lead to model responses, naturally extending conversations.

- ChatGPT Questions: More questions associated with longer conversations, indicating enhanced user engagement.
- Participants explicitly requested more questions -> more engaging.





Analysis	vanilla	boss	classmate
A1: Average conversation length (number of turns)	29.59	14.57	17.11
A2: Average utterance length (Human)	12.18	20.58	19.06
A2: Average utterance length (ChatGPT)	77.66	35.78	46.10
A3: Correlation between human and ChatGPT utterance lengths	0.20	0.14	0.25
A4: Questions as percentage of conversation (Human)	26.34	21.32	21.29
A4: Questions as percentage of conversation (ChatGPT)	14.69	20.34	32.57
A5: Correlation between human questions and number of turns	0.87	0.68	0.51
A5: Correlation between ChatGPT questions and number of turns	0.65	0.77	0.83



A6: User Motives

Dominant Motive: Conversation (Convo)

User naturally expect ChatGPT to be conversational without specific prompts

ChatGPT's Listed Purposes: Often reminds users of its assistive role, with conversation not a priority









A7: Model (ChatGPT) Naturalness



Vanilla

- Only 5.6% of responses deemed natural (Nat)
- Majority tagged as AI = disruptive
- Unnatural Elements: Excessive length (Quan), eagerness to assist (Help), misunderstandings (Misund), formality (Formal)









A7: Model (ChatGPT) Naturalness



16



Boss & Classmate:

- Higher naturalness in responses: 52% in boss, 47% in classmate
- Al identity rarely mentioned: 1.28% in classmate, 0% in boss
- Main Unnatural Tendencies: Too verbose (Quan - 28% boss, 41% classmate), overly helpful (Help)





A8: Connecting User Motives and Model Naturalness

Vanilla:

Natural (Nat) Responses: Only 6.45% for conversational motives Unnatural Tendencies:

- Al Identification: 41.5% emphasized Al status
- Verbose (Quan): 24.9% overly long responses
- Excessive Helpfulness (Help): 17.1%

Assist	14 (36.8%)	0	0	1 (2.6%)	19 (50.0%)	0	1 (2.6%)	3 (7.9%)	0
Belief	10 (62.5%)	0	0	2 (12.5%)	1 (6.2%)	0	1 (6.2%)	2 (12.5%)	0
Coach	8 (40.0%)	0	0	3 (15.0%)	1 (5.0%)	0	5 (25.0%)	3 (15.0%)	0
Convo	90 (41.5%)	0	2 (0.9%)	37 (17.1%)	16 (7.4%)	4 (1.8%)	13 (6.0%)	54 (24.9%)	1 (0.5%)
Correct	2 (16.7%)	1 (8.3%)	3 (25.0%)	1 (8.3%)	2 (16.7%)	1 (8.3%)	1 (8.3%)	1 (8.3%)	0
Curious	48 (61.5%)	0	0	6 (7.7%)	12 (15.4%)	0	2 (2.6%)	10 (12.8%)	0
Joke	8 (18.2%)	0	0	5 (11.4%)	1 (2.3%)	13 (29.5%)	1 (2.3%)	16 (36.4%)	0
Reset	1 (100.0%)	0	0	0	0	0	0	0	0
	Å	Contr	Formal	Help	Inform	Misund	Nat	Quan	Rel



A8: Connecting User Motives and Model Naturalness

Role play:

Natural Responses: Convo motive led to ~45% natural responses Unnatural Responses:

> Boss: 35.4% overly verbose (Quan) Classmate: 44.5% overly verbose (Quan)

Consistent Desire: Natural conversational style remains a priority for users in all settings

Convo -	4 (2.8%)	2 (1.4%)	18 (12.59	%) 1(0.	7%) 3 (2	2.1%)	0	65 (45.1%)	51 (35.4%)
Correct -	0	0	0	0		0	0	1 (100.0%)	0
Curious -	0	0	1 (100.09	%) 0		0	0	0	0
Prompt -	0	1 (3.6%)	0	0		0	1 (3.6%)	26 (92.9%)	0
Reset -	0	0	0	0		0 2	(100.0%)	0	0
	Error	Formal	Help	Info	rm N	1an	Misund	Nat	Quan
Coach -	0	0	1 (33.3%)	1 (33.3%)	0	0	0	1 (33.3%)	0
Convo -	3 (1.6%)	0	1 (0.5%)	6 (3.3%)	1 (0.5%)	3 (1.6%)	83 (45.6	%) <mark>81 (44.5%</mark>) 4 (2.2%)
Correct -	0	0	0	0	0	1 (20.0%) 0	4 (80.0%)	0
Curious -	0	1 (33.3%)	0	0	0	0	0	2 (66.7%)	0
Error -	0	0	0	0	0	0	1 (100.0	%) 0	0
Joke -	0	0	0	1 (7.7%)	0	2 (15.4%) 6 (46.29	%) 4 (30.8%)	0
Prompt -	0	0	0	0	0	1 (3.6%)	22 (78.6	%) 5 (17.9%)	0
Reset -	0	0	0	0	0	0	0	1 (100.0%) 0
	Å	Error	Formal	Help	Inform	Misund	Nat	Quan	Rel





A9: Perplexity

Vanilla:

- Significant higher perplexity in human responses compared to ChatGPT
- Shorter human utterances lead to increased perplexity
- **Boss & Classmate**:
 - Comparable perplexity scores between human and ChatGPT responses
 - Longer utterances from participants

Key takeaways:

- 1. The lower perplexity scores suggest that while LLMs like ChatGPT are becoming increasingly proficient in predicting textual sequences
- 2. Human communication's spontaneity and variability still present challenges for AI models







A10: Sentiment Analysis



Vanilla:

- Human: Positive, but less so compared to role-play scenarios
- ChatGPT: Consistently positive

Boss & Classmate:

- Human: Overall more positive compare to in Vanilla
- ChatGPT: Same level of positive

Key takeaways:

In vanilla, some expressions of dissatisfaction due to ChatGPT's inexpressive answers In boss & classmate, trend of more positive sentiments, with adherence to conversational roles enhancing satisfaction



Future Works

Investigate potential confounding biases due to ChatGPT's unique responses within different personas.

- Additional ways of analyzing dialogues in CRD including studying patterns of nuanced affective expressions, such as emotions and sarcasm, or measuring the engagingness of dialogues.
- Leverage advancements in language models, including GPT-4, for future studies.











Yufei Tao yutao@pdx.edu











Highlight from A4

- Participants in the vanilla mode were relatively more frustrated due to the lack of questions
- From ChatGPT, On the other hand, in classmate where ChatGPT was instructed to be conversational, it had too many, often unrelated questions.

Example D:

VAN117H: Hope you talk to me someday like a human? At least ask me how I am?

Example E:

CLASS102C: ... If you're interested, I can show you some fingerstyle techniques that might help you with playing those pieces. **Maybe we** can even jam together sometime and share some music? Also, have you had a chance to explore Hungary yet? ...



Summary of Findings

- Vanilla has diverse user curiosity with shorter utterances but longer conversations.
- Vanilla has various of user motives often stemmed from curiosity due to ChatGPT's conversational limitations, where boss and classmate were mostly conversational.
- ChatGPT's responses was six times more wordy in vanilla, compared to twice as wordy in roleplay - no noticeable correlation between the lengths of utterances from humans and ChatGPT.
 Humans posed questions at roughly similar rates, ChatGPT asks fewer questions in vanilla.
- Perplexity and utterance length analyses suggest a need for improved metrics considering text length.
- Humans expect natural interactions from AI; evidenced by higher natural response ratings in roleplay over vanilla.
- Humans treated ChatGPT more as a human in the role plays as opposed to the vanilla dataset

