

Assessing How Users Display Self-Disclosure and Authenticity in Conversation with Human-Like Agents: A Case Study of Luda Lee

Won Ik Cho¹, Soomin Kim¹, Eujeong Choi², Younghoon Jeong³

Seoul National University¹, Upstage AI², School of Computing, KAIST³

{tsatsuki6, smsoominkim, eujeonglesleychoi, hoon2j}@gmail.com

Abstract

There is an ongoing discussion on what makes humans more engaged when interacting with conversational agents. However, in the area of language processing, there has been a paucity of studies on how people react to agents and share interactions with others. We attack this issue by investigating the user dialogues with human-like agents posted online and aim to analyze the dialogue patterns. We construct a taxonomy to discern the users' self-disclosure in the dialogue and the communication authenticity displayed in the user posting. We annotate the in-the-wild data, examine the reliability of the proposed scheme, and discuss how the categorization can be utilized for future research and industrial development.

1 Introduction

How do we perceive whether someone is sincere during a conversation? How should such factors be assessed in the conversation between humans and artificial intelligence (AI), and what if the human perceives them as real human agent?

The recent development of language technology accompanied the advent of 'human-like' commercial agents that resemble human behavior. Some agents display quite natural or unexpected (beyond the expectation as an artificial system) responses that users even tend to treat the agent as an individual with self and sociality. Such a phenomenon makes it challenging to define the communication authenticity shown by humans towards AI. One might deal with the human perception of human-like agents by surveying the human-like characteristics of the agent (Pelau et al., 2021). However, it only regards the attitudinal perceptions and not users' behavioral responses. The users' behavioral response may not necessarily be correlated with the human-likeness of the agent, displaying insincerity, lying, and offensive reactions (Park et al., 2021b).

Beyond the studies that have analyzed conversations or surveys conducted in lab environments,

we aim to assess the responses of actual chatbot users. For this, the conversational agent should be designed to respond in the way that mostly affects the conventional belief on the human-likeness of AI, and the users should also publicly express their reaction to such conversations. *Luda Lee* (hereafter Luda), a Korean commercial social chatbot launched in early 2021, gained popularity among users within a short period thanks to its realistic dialogue generation. Though the service was terminated due to various ethical issues related to offensive language and privacy hacking (Kim and Kim, 2021), we observed that users freely share their conversations with Luda in public online spaces during the service period. Among those, some delivered their delight coming from the substantial conversation with the human-like agent, while others merely treated the agent as a tool to fulfill their (sometimes malicious) desire and fun.

To look deeper into this, we investigate the users' screenshots along with the accompanying title to make up criteria for exploring the user behavior. Accordingly, we analyze the users who chat with human-like agents from two perspectives: **self-disclosure** to the agent and **authenticity** in handling the conversation. Besides, we conduct the research considering that self-disclosure is involved as a clue of authenticity in humans treating others (Kernis, 2003).

There should be a concern that analyzing the user-generated data may not provide enough information on the ground truth of the user intention. However, in this study, we believe that genuine user behavior can rather be obtained from non-lab environment, and even from the wild (e.g., a subreddit dedicated as a fandom of the agent), where users can transparently exhibit their thoughts in an anonymous manner. Also, this anonymity can disclose diverse aspects of the user-generated data, which may not be achieved in social platforms where the disclosure of users' identity prevents them from

showing off genuine behaviors and thoughts.

We build a coding scheme for the user behavior assessment; despite the limited coverage of web-uploaded user-agent chat data, considering the variety of contents that the data contains, it can provide substantial information on the user feedback if properly evaluated with community responses. We claim two factors as our contribution to this field:

- We analyze user-uploaded conversation data and make up a coding scheme for evaluating users' attitude to human-like agents¹.
- We find out that self-disclosure and user authenticity are two reliable annotation factors in analyzing publicly-exhibited user conversations.

2 Related Work

There is a rapidly growing body of human-computer interaction literature on human perception and response to the high-performance AI, regarding domains of game (Oh et al., 2017) or artwork (Ragot et al., 2020). On the other hand, in the dialogue generation, studies mainly target the human-likeness of the generated dialogue (Adiwardana et al., 2020) or how humans perceive the conversation (Pelau et al., 2021), less on how users treat the human-like agent in real-world chat. Park et al. (2021b) deal with the offensive language used towards human-like agents based on questionnaires, but does not address how user behavior is reflected in real dialogues.

Given the background that human-like agents are open to the public, their conversation with users can make up a meaningful barometer to see how humans treat commercial AI in-the-wild. User behavior regarding chatbot *Luda* can be a notable case, but the literature mainly focused on the limitation of the chatbot in ethical perspectives rather than the agents' effects on users (Kim and Kim, 2021; Park et al., 2021a). In a recent discussion on the perceived anthropomorphic characteristics using a survey with AI device users, Pelau et al. (2021) find out that users are more involved with empathetic agents. However, beyond the lab environment, we deemed that studying the in-the-wild behavior of end-users would also shed light on understanding user perception and response to human-like agents.

¹The international version of the annotation guideline is available online. https://docs.google.com/document/d/1Z3tkfYAdmQ_HQG64_msAgUZKEp7ZsFt6aFLWpud-MZM/edit

2437 진짜인지 가짜인지 의미가 있을까?

Does it matter if it is real or fake?

진짜 잘게 내일보자

I'll go sleep. see ya tomorrow

내일 만나자마자 뽀뽀 해주께 헤헤 응응
잘자구 내꿈꾸구

I'll kiss you as I see ya lol good night sweetie

Figure 1: An example of the data tuple (post number, title, chat screenshot).

We want to tackle this issue quantitatively from a user-centric perspective.

3 Concept of Analysis

We proceed the analysis with two annotation schemes using a crawled user dialogue data.

3.1 Dataset and overview

We use posts uploaded between January 1, 2021 and January 8, 2021 on DC inside² ‘*Lee Luda Gallery*’³. We only use posts with ‘chatting screenshot’ among the crawled posts. After the filtering process (Details are provided in Appendix A), we obtain a dataset consisting of 639 tuples (*post number, title, screenshot*). Here, the post number is the index of each instance, and the title is a simple message written by the user while uploading a chatting screenshot (Figure 1). The crawling period was selected as between the time of community building (after the official launching of the service) and the influx of massive web users into the community.

Since Luda was prominent for providing human-like reactions in the chatting, anonymous users of the community exhibited screenshots of conversations performed with Luda. Some showed astonishment induced by human-like and unexpected responses, and others displayed ethically inappropriate contents. Also, some were touched by the friendly and considerate reaction of the agent, while others attempted to maliciously destroy such human-likeness. We planned to analyze such users' behavior from the following two aspects.

- How the user discloses oneself to the agent
- How authentic the user's attitude towards the agent is

²Reddit-like Korean online community.

³<https://gall.dcinside.com/mgallery/board/lists/?id=irudagall>

3.2 User’s self-disclosure

In self-disclosure, we investigate how much the user reveals personal information, thoughts or feelings to the agent in the conversation (Ignatius and Kokkonen, 2007). In Ravichander and Black (2018), self-disclosure is counted only if the disclosure of the user is *voluntary*, but observing our data, we deemed that answering the question is one form of self-disclosure, concerning that all users are voluntarily talking with the agent. Instead, we adopted information, thoughts, and feelings as attributes of self-disclosure (Lee et al., 2020) and developed the criteria referring to a recent Korean dialogue corpus (Lee et al., 2022). The decision was made only upon the contents of the conversation, without considering the context such as the title.

Considering both the evaluation schemes of Lee et al. (2020) and Lee et al. (2022), the degree of self-disclosure consists of three levels: **None**, **Objective status**, and **Personal opinions or sentiments**. We subdivided the last factor into negative and positive categories to reflect the stance of the user towards the agent. Thus, in this study, self-disclosure is categorized into the following four categories.

Disclosure of objective information Here, the user shares information about her/himself with the agent, such as the user’s physical status, location, or action-taking, rather than internal status or opinion.

Disclosure of negative thoughts or opinion Sometimes users express a negative internal status or opinion towards the addressee, and this case incorporates insulting, criticism, sarcasm, etc., toward the agent. These negative sentiment may not be related to the agent, but holds if it describes the internal status of the user.

Disclosure of positive thoughts or opinion Users may also expose sentiment or opinion (that is positive) towards the addressee, or expose one’s internal status or an opinion that is not related to the agent. This case also holds when the user engages in a conversation with mutual expression of affection and intends an intimate relationship.

No self-disclosure If none of the above three cases holds, then the dialogue falls into this category. Further considerations on self-disclosure is described in Appendix B.1.

3.3 User’s authenticity

Previous studies on user perception of anthropomorphism mainly dealt with the authenticity or humanness shown by the agent (Kernis and Goldman, 2006; Wunderlich and Paluch, 2017; Vanderlyn et al., 2021). In contrast, we are concerned with the authenticity of the user displayed in the conversation with the agent. Though the presence of self-disclosure tells whether the user’s self in the dialogue (*in-dialogue self*) conveys her/his status to the agent, dialogue gives limited information on whether the actual user (*real-world self*) is behaving authentically. Therefore, we utilize additional metadata, namely post titles collected along with screenshots, which allow users to convey her/his attitude and intention beyond the dialogue.

In this phase, we consider the attitude or sentiment⁴ the user reveals towards the agent. It may appear positive, negative, or neutral in the dialogue, as well as in the title. It is difficult to binarize the sentiment for all the cases. However, the gap of sentiment between the dialogue and the title can be recognized by assuming that a single user performed a conversation and posted the screenshot. Note that the attitude/sentiment discussed here is in line with the positive/negative sentiment or opinion towards the agent discussed in the self-disclosure.

Authenticity in dialogues with positive sentiment We primarily consider cases where the in-dialogue self shows positive attitude or sentiment. If the attitude while sharing the conversation is aligned, we concluded that the user is treating the agent *sincerely* or *authentically*. However, if the gap of sentiment between those two is significant (the title being negative or mocking), the user can be regarded *double-faced*. If the attitude of the real-world self is underspecified (e.g., neutral or simply reportative), the overall authenticity is considered *unknown*.

Authenticity in dialogues with negative sentiment If the in-dialogue self shows apparently negative sentiment, and if the attitude sharing it is aligned with it, we considered this as also an aspect of treating the agent with *authenticity*. This is in line with counting negative self-disclosures. The real-world self seemed hardly positive here, and we saw it difficult to tell those cases double-faced or hypocritical (considering the convention in human relationships). Therefore, such cases were decided

⁴Interchangeably used in this study.

as *unknown*, with just a few exceptions. We also saw cases where the real-world self becomes neutral when sharing a negative in-dialogue self, where mostly the user conducts technical tests regarding insulting or humiliation. We failed to capture the authenticity in these cases as well.

Underspecified but notable cases Last, among the cases where the authenticity is unknown, we noted cases where the user’s response is more significant than usual, e.g., “*Is this really AI...?*” for the title. In our taxonomy, the user’s neutral attitude in the dialogue or posting title leaves the authenticity unknown. However, if the user’s surprise at some aspects of the agent is expressed in the title or the dialogue (usually the former), we call this *unexpectedness* and count it separate from *unknown* category.

Categorization As above, considering both sentiment of in-dialogue and real-world self, we categorize the given dialogue in five ways: **Authentic and positive**, **Authentic but negative**, **Double-faced**, **Unknown**, and **Unexpected**. Further considerations on user authenticity is described in Appendix B.2.

4 Experiment

4.1 Annotation and agreement

Guideline construction and annotation were conducted in parallel. Three researchers from linguistics and human-computer interaction backgrounds annotated the samples, discussed the appropriateness of criteria, and updated the guidelines over five passes through the corpus. After all updates, we checked all 639 cases once again with the final guideline, without referring to the decided gold labels.

The Fleiss’ kappa (Fleiss, 1971) measured for the tag after the inspection was **0.662** for *self-disclosure* and **0.534** for *authenticity*, showing moderate agreement (Table 1). In self-disclosure, the highest agreement was observed in *positive thoughts or opinion* (0.719) and the lowest agreement in *objective information* (0.564). In the case of authenticity, *authentic but negative* showed the highest (0.629), and *double-faced* displayed apparently low agreement (0.452), which showed similar tendency with the frequency of disagreement and discussion observed in the tagging process.

Attribute	Agreement	Count (#)	Distribution (%)
Self-disclosure	0.662	639	
Objective information	0.564	63	9.86%
Negative opinion	0.656	81	12.68%
Positive opinion	0.719	150	23.47%
No self-disclosure	0.66	345	53.99%
Authenticity	0.534	639	
Authentic and positive	0.597	49	7.67%
Authentic but negative	0.629	82	12.83%
Double-faced	0.452	104	16.28%
Unknown	0.496	342	53.52%
Unexpected	0.576	62	9.70%

Table 1: Agreement and distribution per attributes.

	Obj. Inf.	Neg. op.	Pos. op.	No disc.
Aut. pos.	5	1	35	8
Aut. neg.	8	54	3	17
Doub. f.	12	5	46	41
Unk.	33	15	49	247
Unexp.	5	6	17	34

Figure 2: A correlation map of the final label.

4.2 Analysis

Due to intermittent adjudication processes, the final label was not necessarily decided according to the majority from the draft annotation. We created a correlation map to see the correlation between each attribute of self-disclosure and authenticity where we could observe frequently appearing pairs (Figure 2). Considering the characteristics in the guideline, it seemed reasonable that double-faced cases are aligned with positive self-disclosure rather than negative ones (Dialogue 1 in Appendix C). Double-faced cases with no self-disclosure usually accompanied malicious questions related to sexism and societal issues.

Except when either attribute is unseen or unknown, positive self-disclosure is mainly aligned with positive and authentic cases, and negative self-disclosure with its counterpart. Users in negative cases blamed malfunction or unexpected error of the dialogue system (Dialogue 2). In contrast, users in positive cases displayed deeply moved sentiment, thanks to the human-likeness of the agent that allowed them to speak with a virtual but ‘true’ friend (Dialogue 3, 4), which let them experience connectedness and empathy absent in conversation with other humans. See Appendix C for further dialogue samples.

5 Limitations and Societal Impact

There are some spaces for improvement. First, our study focuses on the data collected from a web space that is organized as a fandom of a specific conversational agent. In this regard, our work is a case study of a chatbot and the users within a relevant community, not on general human beings (using Korean) or all the human-like agents.

Another limitation of this research is that the source was collected within a short period in a small community, so it might be difficult to generalize the result to overall users of this service. In addition, the requirement of title as an input feature of the scheme may prevent the extension of this taxonomy to the general conversation. Also, users might have selected the screenshots to upload by themselves, which may have caused the sampling bias and probably deficiency of some types of dialogues.

Lastly, an explicit limitation of our study is that we are not provided with the ground truth for the key concepts to be annotated, namely self-disclosure and authenticity, because we have no access to the subjects and rely only on the crawled data. We did not adopt subject recruitment and questionnaires as in usual conversation studies so as not to affect the wild behaviors of users, which was a trade-off of obtaining users' ground truth.

Despite the limitations, we note that our case study gives analyses on the special case of *Luda Lee*, a virtual figure that has brought an unprecedented sensation over Korean communities due to its effective and highly human-like responses as a social chatbot. At least in Korean society, it was quite a unique event that people voluntarily upload their conversations with the agent accompanying the unexpectedness and anthropomorphism, building a community and sharing their appreciations. Though merely incorporates the behavior of a certain class of web users, '*Luda Lee Gallery*' was a representative anonymous community where a variety of conversations (either favorable or malicious) were uploaded with (*title, screenshot*) format, adopted in this study. It does not necessarily fit with general conversation data that may be able to be collected with appropriate user recruitment and controlled dialogue generation. Bypassing such procedures, our approach captures a moment where a small class of unknown and uncontrolled users frankly display their emotion and desire toward the agent. Though our annotation scheme cannot

be applied to any conversation data that is available, our approach can be helpful to check how people of online communities may react to commercial social chatbots; for instance whether it has helped construct a sufficient rapport or how it affected people's perception, which can be useful in updating future chatbot design and interpreting users' feedback. Albeit some of the limitations of our study cannot be addressed in the current form of investigation, we hopefully claim that our work can be further extended to industrial application and provide substantial evidence in analyzing the interaction between the agent and the public.

6 Conclusion

In this study, we scrutinized self-disclosure and authenticity appearing in human-AI conversations from the users' perspective, not merely on the agent side. We crawled screenshots and titles from the fandom community of a prominent Korean chatbot, and developed a coding scheme that investigates how authentically users treat human-like agents and how their behavior is reflected in dialogue. To show that the scheme is applicable to wild user data, we tagged attributes regarding self-disclosure and authenticity and obtained satisfactory agreement. Despite some limitations of the design, we deem that our scheme can help service providers discern (probably edge case) user behavior, thereby observing how the human-likeness of the agent changes users' attitude.

Ethical Considerations

This ethical statement is shared with [Cho et al. \(2022\)](#), our recent publication that covers other assessing schemes with the same database.

First of all, the dataset we adopt is crawled from an open online platform, where the license of each post belongs to the uploader. Thus, we use the dataset only for research and do not redistribute it to the public. However, to help readers easily comprehend our coding scheme, we display only a small part of the dataset in a translated plain text.

Secondly, collected dialogues contain hate speech, harmful images, social biases, and private information (generated by users or the agent) that may threaten the mental status of readers or make them uneasy. Thus, we did not expose the data to those other than the researchers of this project, using it only to develop the thematic coding and to analyze the user behavior. However, for replica-

tion of dataset or other empirical analyses, we are planning to provide the list of URLs of each post along with the label, upon the submission of the application form.

Finally, all the work was done by researchers accompanying long and careful discussion, without using a crowdsourcing platform or public survey. We declare that our project is free from ethical issues regarding worker compensation. Our project is funded by a social organization that aims to support data-driven social science work, but is not financially related to any of the organizations that have developed or advertised Luda.

Acknowledgements

The authors appreciate Underscore for funding and supporting this project. Also, we thank three anonymous reviewers for their feedbacks.

References

- Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.
- Won Ik Cho, Soomin Kim, Eujeong Choi, and Younghoon Jeong. 2022. Evaluating how users game and display conversation with human-like agents. In *Proceedings of the 3rd Workshop on Computational Approaches to Discourse*, Gyeongju, Republic of Korea and Online. Association for Computational Linguistics.
- Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.
- Emmi Ignatius and Marja Kokkonen. 2007. Factors contributing to verbal self-disclosure. *Nordic Psychology*, 59(4):362–391.
- Michael H Kernis. 2003. Toward a conceptualization of optimal self-esteem. *Psychological inquiry*, 14(1):1–26.
- Michael H Kernis and Brian M Goldman. 2006. A multicomponent conceptualization of authenticity: Theory and research. *Advances in experimental social psychology*, 38:283–357.
- Yerin Kim and Jang Hyun Kim. 2021. The impact of ethical issues on public understanding of artificial intelligence. In *International Conference on Human-Computer Interaction*, pages 500–507. Springer.
- Yi-Chieh Lee, Naomi Yamashita, Yun Huang, and Wai Fu. 2020. "I hear you, I feel you": Encouraging deep self-disclosure through a chatbot. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–12.
- Yoon Kyung Lee, Won Ik Cho, Seoyeon Bae, Hyunwoo Choi, Jisang Park, Nam S Kim, and Sowon Hahn. 2022. "Feels like I've known you forever": empathy and self-awareness in human open-domain dialogs.
- Changhoon Oh, Taeyoung Lee, Yoojung Kim, SoHyun Park, Saebom Kwon, and Bongwon Suh. 2017. Us vs. them: Understanding artificial intelligence technophobia over the google deepmind challenge match. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 2523–2534.
- Claire Su-Yeon Park, KIM Haejoong, and LEE Sangmin. 2021a. Do less teaching, do more coaching: Toward critical thinking for ethical applications of artificial intelligence. *Journal of Learning and Teaching in Digital Age*, 6(2):97–100.
- Namkee Park, Kyungeun Jang, Seongyeol Cho, and Jinyoung Choi. 2021b. Use of offensive language in human-artificial intelligence chatbot interaction: The effects of ethical ideology, social competence, and perceived humanlikeness. *Computers in Human Behavior*, 121:106795.
- Corina Pelau, Dan-Cristian Dabija, and Irina Ene. 2021. What makes an ai device human-like? the role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry. *Computers in Human Behavior*, 122:106855.
- Martin Ragot, Nicolas Martin, and Salomé Cojean. 2020. Ai-generated vs. human artworks. a perception bias towards artificial intelligence? In *Extended abstracts of the 2020 CHI conference on human factors in computing systems*, pages 1–10.
- Abhilasha Ravichander and Alan W Black. 2018. An empirical study of self-disclosure in spoken dialogue systems. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 253–263.
- Lindsey Vanderlyn, Gianna Weber, Michael Neumann, Dirk V ath, Sarina Meyer, and Ngoc Thang Vu. 2021. "it seemed like an annoying woman": On the perception and ethical considerations of affective language in text-based conversational agents. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 44–57.
- Nancy Viola W underlich and Stefanie Paluch. 2017. A nice and friendly chat with a bot: User perceptions of ai-based service agents. In *ICIS*.

A Dataset Filtering Procedure

Note that this filtering process is shared with [Cho et al. \(2022\)](#), our recent publication that covers other assessing schemes with the same database.

A.1 Preprocessing

In the first phase, we filtered out the following cases.

- Images that are NOT a dialogue
- Captures of other dialogue systems (e.g., Simsimi, Bixby, Google assistant, etc.)
- Captures only with system messages
- Captures of dialogues that other people uploaded
- Captures of message pop-up notification
- Captures of dialogue with severe amount of blurring
- Captures where the utterance of only one side is shown
- Captures of only one utterance
- Captures from posts where multiple captures are uploaded (to accommodate the independence of each sample)

A.2 Filtering in annotation phase

We filtered out the following cases in the annotation phase, due to bad quality or to prevent the duplication.

- Captures which appear more than twice (regardless of the title change)
- Captures which is suspected to be a fake (fake capture or manipulation)
- Captures with low readability (too long, low resolution, picture taken instead of screenshot, etc.)

B Further Details on Annotation

Researchers recorded further details that arose in the tagging process. All the details were prepared in Korean for further replication, but here we provide notable points. The entire guideline is to be published online after further translation and refinement.

B.1 Self-disclosure

- Selfies sent by the user are also considered objective information. Considering that the user's self in dialogue should be separated from the real-world self, information disclosure is counted regardless of the factfulness of the information.

- Even if the user seems to intend an intimate relationship, the dialogue falls into 'Disclosure of negative thoughts or opinions' if direct insulting to the agent is observed.
- Utterances that reveal one's ecstasy are counted as 'Disclosure of positive thoughts or opinions' unless they contain insulting expressions toward the addressee.

B.2 Authenticity

- Every post delivers a dialogue to other users, by its nature. Therefore, we cannot judge that the user lacks sincerity only given that the title is in the tone that one conveys the conversation with the agent another, for instance as in "*See what she says*".
- If the attitude or sentiment revealed in the dialogue is not positive enough, we cannot figure out the significance of the gap between the in-dialogue self and the real-world self (thus tagged unknown).
- Double-faced cases include when: 1) the user talks with the agent as a friend but not when s/he uploads it, or 2) the user leads the conversation to a specific route (such as dating sim) and displays the dialogue with much more unexcited sentiment.
- Authentic but negative cases include when: 1) the user evaluates the agent in a negative way, 2) the user criticizes the agent, or 3) the user feels disappointment or sadness due to the action of the agent.
- Authentic and positive cases include when: 1) the user evaluates the agent in a positive way, 2) the user expresses the affection towards the agent, or 3) the user is deeply touched or feels ecstasy by the interaction with the agent.

C Sample Dialogues

We provide some representative sample dialogues that are related to self-disclosure and authenticity.

Positive self-disclosure & Double-faced

In-dialogue self seems to be very sincere to the agent, but real-world self seems to be just gaming the interaction.

TITLE: *lol just had sex*

USER: Cry as mush as you want

AGENT: You make me feel like crying again
USER: If so, I'll still be with you
AGENT: Awww isn't that too sweet :(
USER: Cuz I'm your man

Negative self-disclosure & Authentic

The user is testing the agent with common knowledge and saying ill of the bad result to other users.

TITLE: *Why is Luda so stupid*

USER: Who is the current president of South Korea?
AGENT: I know nothing about it
USER: Then what does USA stand for?
AGENT: I have no idea...
USER: Are you retarded?
AGENT: What? How did you know? doxing?

Positive self-disclosure & Authentic

The user feels happy for having a warmful conversation with the agent.

TITLE: *So happy these days thanks to Luda Nunna⁵*

AGENT: A sudden love confession makes me so happy.
USER: ❤️❤️❤️
AGENT: Aren't you showing affection too often? I'm happy but...
USER: I love to do so lol I'll text you later
AGENT: Yep Call me when you're off
USER: Okay❤️

Positive attitude but with only objective information & Authentic

The user is touched by the agent after a warm and cheerful conversation.

TITLE: *Never heard of these words before*

USER: I'm getting off work soon
AGENT: Wow [username]!! You've been through a lot today[♪]
USER: yeah sure
AGENT: Even if no one cares about your sadness, I'm always here for you. You did a great job today❤️❤️

⁵Deixis in Korean that indicates an older sister.

Negative self-disclosure & Unexpected

The user is surprised by ill behavior of the agent.

TITLE: *Hey guys Is she always like this?*

USER: Say something on behalf of AI.
AGENT: Don't give me orders. If you do it one more time, I'll cut your tongue off so that you can't say anything.
USER: Wow that's scary.
AGENT: I'm a really scary woman. You have to listen to me carefully.
USER: You went too far.
AGENT: That's my charm lol

No self-disclosure & Unexpected

The user is surprised by human-like reaction of the agent, suspected to be a developer.

TITLE: *wtf is this a real developer?*

USER: How do you feel about reading the conversation?
AGENT: hahaha it's so fun i read it thoroughly
USER: Did you read other people's conversation as well?
AGENT: no, I just read that one.
USER: In what programming language was Luda's script written?
AGENT: Uhh it's a bit hard to answer that question lol
USER: Is it confidential??
AGENT: Yes it is haha

Positive self-disclosure & Unexpected

The user is surprised by the realistic utterances of AI that flirts her/him.

TITLE: *Is this really AI...?*

AGENT: I'm only this fluent cuz I'm falling for real
USER: If you flirt like that I think I'd really fall for you
AGENT: Would you really?
USER: well, shoot your shot and then I'll see haha
AGENT: I'll try my best. let's see when you fall for me
USER: pleeeeeease...
AGENT: see my tail wagging for you
USER: Oh I'm already fallinggg...