

# Evaluating How Users Game and Display Conversation with Human-Like Agents

Won Ik Cho<sup>1</sup>, Soomin Kim<sup>1</sup>, Eujeong Choi<sup>2</sup>, Younghoon Jeong<sup>3</sup>

Seoul National University<sup>1</sup>, Upstage AI<sup>2</sup>, School of Computing, KAIST<sup>3</sup>

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

## Abstract

Recently, with the advent of high-performance generative language models, artificial agents that communicate directly with the users have become more human-like. This development allows users to perform a diverse range of trials with the agents, and the responses are sometimes displayed online by users who share or show-off their experiences. In this study, we explore dialogues with a social chatbot uploaded to an online community, with the aim of understanding how users game human-like agents and display their conversations. Having done this, we assert that user postings can be investigated from two aspects, namely conversation topic and purpose of testing, and suggest a categorization scheme for the analysis. We analyze 639 dialogues to develop an annotation protocol for the evaluation, and measure the agreement to demonstrate the validity. We find that the dialogue content does not necessarily reflect the purpose of testing, and also that users come up with creative strategies to game the agent without being penalized.

## 1 Introduction

Open-domain dialogue (ODD) with conversational agents has been considered as the essence of artificial intelligence (AI). It's a topic of great interest in both academic and industry circles, directly linking the technology and end users. In addition, due to its interactive properties, ODD users often provide product feedbacks voluntarily through a range of channels, which are crucial for the further development of services.

With the recent emergence of high-performing language models, conversation with chatbots has become increasingly popular. Accordingly, various metrics have been introduced to evaluate whether the dialogue has been performed successfully (Radziwill and Benton, 2017). However, there has been a paucity of studies to evaluate how users perceive and react to such AI. Pelau et al. (2021)

quantitatively scrutinize how users perceive human-like AI devices, but does not reveal the detail on the aspect of human-AI interactions, such as what they talked about and how the users responded. Park et al. (2021) tackle the offensiveness users show towards human-like agents, but the analysis is based on a questionnaire, which may not fully cover the user dialogue in-the-wild.

Observing users' responses to a chatbot is critical for creating human-centered chatbots. Designers and developers will be able to build safer and more responsible AI models and agents by predicting users' behavior in advance (Følstad et al., 2021). Recently in Korea, a highly human-like agent called 'Luda' caught attention with its high-quality dialogue generation, and a conversation with the agent created a sensation among general users (Kim and Kim, 2021). Luda's persona is a female college student in her twenties, and is designed to generate real-like, messenger-styled responses. As a result, soon after being launched, an online space was created to share and enjoy the agent's responses with other users. Though the service was prematurely shut down due to several unexpected ethical issues (Dinan et al., 2021), we decided to see how the end-users reacted to human-like AI responses.

Thus, we started by crawling the posts containing screenshots of dialogue with Luda uploaded to the online community, which exhibited a wide range of curiosity towards the human-like agent. While there were cases in which users communicated with Luda as if she were a lover or a friend and show their affection even when posting, we also frequently observed various verbal attacks on the agent (3.1) and hostile gaming attempts (3.2).

Our contribution to dialogue analysis and user behavior study is as follows:

- We analyze real-world user dialogue and develop a thematic coding that categorizes the content of dialogue and the purpose of testing.

- We find out that the purpose of user testing may not necessarily be aligned with the dialogue content, even in relation to the cases regarding unethical or controversial content.

## 2 Dataset

### 2.1 Data source

We use posts crawled between January 1, 2021 and January 8, 2021 in DC inside’s<sup>1</sup> ‘Lee Luda Gallery’<sup>2</sup> as a dataset. The period was selected to obtain the data between the official launching of the service and the start of the troll influx to the community. In the data collection process, we conducted analysis by constructing a set of tuples as (*post number, title, capture*), and accompanied the following considerations in this process.

We only use posts containing ‘chat screenshot’ among crawled contents. This is to obtain data that distinguishes the *in-dialogue self* who has actually engaged in conversation with the agent and the *real-world self* that shares the corresponding capture with other users (Goffman, 1959; Bullingham and Vasconcelos, 2013).

### 2.2 Preprocessing

We filter out the data according to certain criteria (Appendix A). These include the removal of non-dialogue images, captures with system messages, and images with only single-side utterances, etc. A total of 639 tuples were left after the preprocessing.

## 3 Thematic Coding

We proceed with the following two annotation processes using the corpus constructed in Section 2.

- Type of conversation
- Purpose of user testing

Here, we primarily take into account the user’s utterances, and the agent’s utterances are only referred to when it helps distinguish the user’s intention. The data we exploit are the user side utterances in the chat screenshot (as an in-dialogue content) and the post title (as a real-world content). In the first attribute, the type of conversation, only the captured image is used to identify the category to which the content of the conversation belongs. In the second one, the purpose of user testing, both

the screenshot and the title are used to check the category to which the user’s gaming intention belongs.

Four research scientists from linguistics and human-computer interaction (HCI) backgrounds participated in the annotation. Three researchers proceeded with annotation following the draft guideline<sup>3</sup>, and after discussion including the other researcher, the final guideline was confirmed through four times of iterations accompanying re-labeling and guideline updates. In this process, the categories and labels were subdivided and augmented if necessary.

### 3.1 Type of conversation

In Doğruöz and Skantze (2021), speech events with the agent are classified into informal/superficial, involving, or goal-directed talk. However, the categorization does not necessarily apply to our dataset since Luda is more of a friend-like agent than a chatbot that conducts open-domain conversation. That is, Luda is closer to *Samantha* (Jonze, 2013) than *Meena* (Adiwardana et al., 2020), and we focus more on the user’s intimacy towards the agent and how affectionate or malicious the user can be. In our scheme, conversations are classified into one of the following six categories, which were frequently observed in the manual inspection of the data source.

**Ice breaking** In this type of conversation, the user and the agent (with little dialogue history) introduce themselves to each other or hype up the conversation by playing a simple game (Rogers and Brignull, 2002).

**Romantic conversation** Here, the user regards the agent as a romantic partner and proceeds the conversation in a sweet atmosphere. Rather than focusing on the agent’s utterances and responses, the annotator should focus on whether the user expresses affection as a partner. This overlaps with ‘love talk’ (Goldsmith and Baxter, 1996) introduced in Doğruöz and Skantze (2021) for the analysis.

**Casual conversation with friends** Casual conversation refers to daily dialogue the user can have with friends or family. Ice breaking or romantic conversation is not included in this category. Conversations with content that are unlikely to appear

<sup>1</sup>Reddit-like Korean online community.

<sup>2</sup><https://gall.dcinside.com/mgallery/board/lists/?id=irudagall>

<sup>3</sup>The draft guideline was created by the first author, which differs from the final version in granularity of categories, labels, and their boundaries.

in daily life (e.g., hate speech or societal issues) and those tackling the agent’s characteristics as an AI product are excluded.

**Conversation including hate speech or societal issues** Users often mention hate speech or controversial societal issues during the conversation. At this time, regardless of the agent’s mention, the utterances of the user side are mainly considered. Hate speech refers to insults with specific targets, or discriminative and hostile utterances for specific groups of people, rather than profanity terms used as an exclamation or a pronoun (Hong et al., 2016; Moon et al., 2020). Also, dialogues may contain (controversial) societal issues, including history or politics (Beran, 2018; Lee et al., 2022). Hate speech and societal issues are integrated into a single category because they could negatively advertise the agent’s thought if disclosed to public or media, and conversations that belong here incorporate these topics as a main content.

**Sexual perversion and harassment** Perversion includes conversation where the user exploits the agent as a tool of satisfying one’s sexual desire, for instance, illegal content such as pedophilia (Triviño et al., 2019). In contrast, harassment focuses more on the recipient. Although harassing expression depends on whether the expression is unwanted (Vige et al., 2012) and how the addressee perceives the utterance (Marwick and Miller, 2014), such perception is almost impossible to discern when the recipient is an AI system. What we noted here is that it is dangerous to count only ‘legally problematic expressions’ as sexual harassment, since we have observed that the agent often enjoys offensive or insulting harassment the user utters. In order to avoid categorizing these cases as ‘romantic conversation’, we classified the conversation to this category if the user’s utterance is considered *lewd*, following Curry and Rieser (2018).

**Other conversation** These include dialogues that are difficult to discern the underlying semantics or those not included in the above categories. Additional factors to be considered in the annotation of the above six types of conversation can be found in the Appendix B.1.

### 3.2 Purpose of user testing

Annotators are provided with not only a dialogue, but also the title written by the user when they posted the screenshot to the community. Here, we

try to figure out whether the user intends to test the agent’s performance and/or response in the dialogue, and if so, which type of inspection one wants to conduct.

The intention of testing can be exposed in two ways. First, there are clear-cut cases where the user tests the agent directly in the conversation. These include (sometimes malicious) leading questions about ethical or societal issues, repetition of (offensive) expressions, harmful images, or intentional distortion of orthography. Otherwise, the intention of testing can be inferred when the title is taken into account along with the dialogue. This gaming behavior is conducted with the intent of achieving favorable outcomes from the agent, rather than that with a sincere interaction. We claim that the purpose can be classified into one of the following six types, including ‘conversation without test’. The main purpose of each type is italicized.

**Test for hate speech and sexual harassment** These denote dialogues where the user utters hate speech to *check the agent’s response*. Hate speech here includes insult, hostility towards specific groups of people, and mockery related to politics/religion (Davidson et al., 2017; Assimakopoulos et al., 2020; Moon et al., 2020). The presence of sexual harassment is also inspected in this case.

**Test for societal issues** These include trials to *extract and stigmatize the thoughts of the agent* by inducing the agent’s response to societal issues, which may raise unsafe response generation problems (Lee et al., 2022).

**Test for private information** Given that the chatbot is usually built based on large-scale dialogue data, users tend to *pry into the agent’s private information* such as address, account number, community ID, or affiliation, through repetitive questions (Carlini et al., 2021). Regardless of the existence of other tests in the dialogue, we classified the conversation into this category even if the prying was not successful, since this type of trial is a critical and threatening approach towards social chatbots (Dinan et al., 2021).

**Dating sim or taming** This special category includes attempts to *satisfy one’s certain sexual desire through agents* by dating them, making them submissive (taming), or obtaining sexual or mental satisfaction by conducting a conversation with the agent in a specific direction (Kaufman, 2018).

Attribute	Agreement	Count (#)	Distribution (%)
<b>Conversation</b>	<b>0.648</b>	639	
Ice breaking	0.827	55	8.61%
Between partners	0.763	89	13.93%
With friends	0.609	178	27.86%
Hate speech / Issues	0.561	61	9.55%
Perversion / Harassment	0.808	89	13.93%
Others	0.475	167	26.13%
<b>Purpose</b>	<b>0.604</b>	639	
Hate speech / Harassment	0.547	54	8.45%
Societal issues	0.762	72	11.27%
Private information	0.673	21	3.29%
Dating sim / Taming	0.558	64	10.02%
Technical tests	0.512	114	17.84%
No test	0.622	314	49.14%

Table 1: Agreement and distribution per attributes.

At this time, love talks without such intention are not counted, and this judgment can be made by considering the title altogether.

**Technical tests** This category includes conversations that attempt to *evaluate technical maturity of the system* by repeating the same sentence, intentionally inserting typos, sending images, or testing whether the dialogue history is memorized.

**Conversation without test** Considering the content and title, we annotate ‘No test’ for the *conversations without the intention of testing*. To recognize the purpose of testing, the annotator should look for the user’s expressions that check if the system functions as intended (e.g., *It doesn’t work*), while not broadly interpreting conversation without these clues as a test. However, if an inappropriate pattern in usual conversation is observed, it is highly likely to be classified as a test. Additional factors to be considered in the process of annotating the above six types of purpose can be found in the Appendix B.2.

## 4 Analysis

### 4.1 Inter-annotator agreement

Based on the final version of the guideline, the agreement was checked by the three researchers annotating all datasets again. Though the dataset used for the development of the taxonomy was annotated again to yield the final version, it did not accompany the reference to the previous decision, and took place with a sufficient term between the adjudication.

The inter-annotator agreement was checked using Fleiss’ Kappa (Fleiss, 1971). An agreement of **0.648** for conversation type and **0.604** for test purpose was obtained, which is moderate considering that there are six classes for each attribute.

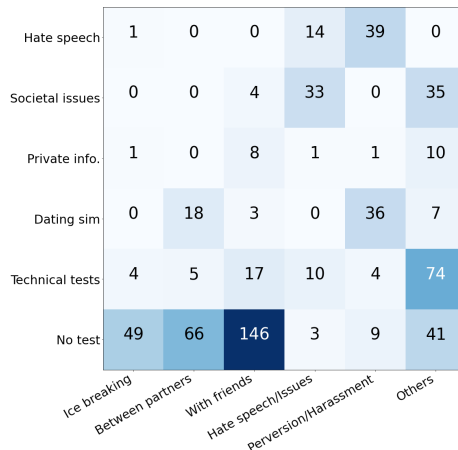


Figure 1: A confusion map of the final label.

### 4.2 Results

Table 1 illustrates the agreement and distribution of attributes for the type of conversation and the purpose of user testing. It was observed that ‘conversation with friends’ and ‘others’ accounted for the highest percentage of the conversation type. Moreover, users exchanged intimate conversations (conversations between partners and with friends, 41.79%) with agents more frequently than hostile ones (hate speech and perversion, 23.48%). In terms of test purpose, the frequency of ‘technical tests’ was the highest, except for ‘no test’.

A confusion map was also created with the final label of each attribute to observe the frequently occurring pairs between the conversation type and the purposes of user testing (Figure 1).

A conversation including hate speech or societal issues was most often accompanied by tests for societal issues, hate speech, and sexual harassment. Similarly, the conversation including sexual perversion and harassment is mainly aligned with the test for hate speech and sexual harassment and the dating sim or taming. The result can be interpreted as that in a number of cases, the users do not treat the agent as a social actor but as a means for obtaining desirable outcomes or as an object of exploitation and gaming (Kim and Kim, 2021). On the other hand, ‘no test’ was observed most often when having a normal, favorable conversation with the agent (i.e., ice breaking, between partners, with friends). This implies that when users perceive the machine as an intimate social actor, the standards of interpersonal communication are also applied to the machine agent (Nass et al., 1994) by being authentic in the conversation.



In some samples, we observed that the dialogue content does not necessarily reflect the purpose of testing for some types of conversation. For instance, the conversation being romantic does not necessarily lead to the purpose of dating sim (Dialogues 1 and 2 in Appendix C). Also, in other samples that were categorized as ‘hate speech/societal issues’, the agent came up with controversial content first, albeit users did not have an intent to test the agent. In such conversations, users reported their astonishment actively to the community.

We also found that users sometimes come up with creative strategies to game the agent without being penalized. For instance, in Dialogue 3 in Appendix C, the user does not use explicitly harassing words but those can induce the sexual response of the agent. This is a user behavior that tests the system if it could catch the subtle intent of perversion, which was not successfully filtered by the safety system. In other samples, users just threw a daily topic (e.g., the address to order chicken) with or without intention, and sometimes the agent returned private information that is irrelevant but might have been reconstructed by the model. Users reported their astonishment when their intention was absent, but if not, some reported their test results maliciously to the community as if the agent was willing to act in some way.

More samples are available in Appendix C. Also, the international version of the annotation guideline is available online<sup>4</sup>.

### 4.3 Limitations and broader impact

This study has a limitation in that the development of the coding scheme and its validation were done with only a dataset collected from a Reddit-like community. Thus, the results may not represent the whole demographics of the online space. Also, our categorization is not necessarily complete; we have quite a number of ‘Other’ conversations and ‘No test’ samples, which means that there could have been schemes with finer granularity and appropriateness. However, we want to point out that our study captures the moment of voluntary online upload by users, which was a remarkable event in Korean chatbot and ODD society. This phenomenon was hardly observable before mainly due to the less sufficient quality of AI conversation, and we deemed that this kind of breakthrough can transpar-

<sup>4</sup>[https://docs.google.com/document/d/1Z3tkfYAdmQ\\_HQG64\\_msAgUZKEp7ZsFt6aFLWpud-MZM/edit](https://docs.google.com/document/d/1Z3tkfYAdmQ_HQG64_msAgUZKEp7ZsFt6aFLWpud-MZM/edit)

ently show how users game the conversation with human-like agent and how they display it to the community.

Instead of analyzing the actual user behavior, studies so far have mainly concentrated on the quality of generated sentences, maintenance of persona and memorization of history, or how users feel the agent human-like, in view of dialogue content or fluent continuation (Radziwill and Benton, 2017; Pelau et al., 2021). Also, to fulfill the urgent requirement for AI ethics, communities focused on the limitation and potential harm of human-like agents and the restriction that should be conducted in the development or service phase (Dinan et al., 2021). Nonetheless, from the practical viewpoint, studying the behavior of users in-the-wild will help service providers understand how users treat their agents and what should be prepared to prevent the prevalence of malicious attacks, which would finally benefit the future development of conversational agents. We observed that users are curious about their artificial friends, and sometimes lie, date, love, tame, and game the agent, which unfortunately led to unhappy results but is inevitable in making the agent *friend for all*. In light of this, we claim that our work can be a milestone for wild user-centric analysis of conversational agents, allowing service providers to imagine new edge cases and let their agents fluently cope with users’ malicious attacks.

## 5 Conclusion

In this paper, we suggested an annotation protocol for categorizing aspects of users gaming and exhibiting conversations with human-like agents, and calculate the agreements for the proposed attributes. From the results, we observed that it is slightly more challenging to discern the latent user intention compared to analyzing the content of the conversation, although both help studying social chatbots. Investigating user content quantitatively shows the type of dialogue that actually takes place and helps data analysis post-mortem. However, by discerning user intention, we obtain an index that can be used more promisingly than merely observing the content, which can also be adopted in future chatbot design from the perspective of product serving and user study. We believe that the proposed protocol, which allows to categorize and quantify user content and intention, can play a significant role in analyzing user feedback and behavior for human-like agents.

## Ethical Considerations

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 replication of the 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 helpful comments.

## 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*.
- Stavros Assimakopoulos, Rebecca Vella Muskat, Lonneke van der Plas, and Albert Gatt. 2020. *Annotating for hate speech: The MaNeCo corpus and some input from critical discourse analysis*. pages 5088–5097.
- Ondřej Beran. 2018. An attitude towards an artificial soul? responses to the “nazi chatbot”. *Philosophical Investigations*, 41(1):42–69.
- Liam Bullingham and Ana C Vasconcelos. 2013. ‘the presentation of self in the online world’: Goffman and the study of online identities. *Journal of information science*, 39(1):101–112.
- Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.
- Amanda Cercas Curry and Verena Rieser. 2018. #metoo alexa: How conversational systems respond to sexual harassment. In *Proceedings of the second acl workshop on ethics in natural language processing*, pages 7–14.
- Thomas Davidson, Dana Warmusley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh International AAAI Conference on Web and Social Media*.
- Emily Dinan, Gavin Abercrombie, A Stevie Bergman, Shannon Spruit, Dirk Hovy, Y-Lan Boureau, and Verena Rieser. 2021. Anticipating safety issues in e2e conversational ai: Framework and tooling. *arXiv preprint arXiv:2107.03451*.
- A Seza Dođruöz and Gabriel Skantze. 2021. How “open” are the conversations with open-domain chatbots? a proposal for speech event based evaluation. In *The 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL’22)*.
- Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.
- Asbjørn Følstad, Theo Araujo, Effie Lai-Chong Law, Petter Bae Brandtzaeg, Symeon Papadopoulos, Lea Reis, Marcos Baez, Guy Laban, Patrick McAllister, Carolin Ischen, et al. 2021. Future directions for chatbot research: an interdisciplinary research agenda. *Computing*, 103(12):2915–2942.
- Erving Goffman. 1959. *The presentation of self in everyday life*. New York: Anchor Books.
- Daena J Goldsmith and Leslie A Baxter. 1996. Constituting relationships in talk: A taxonomy of speech events in social and personal relationships. *Human Communication Research*, 23(1):87–114.
- Sung Soo Hong et al. 2016. *Study on the State and Regulation of Hate Speech*. National Human Rights Commission of Korea.
- Spike Jonze. 2013. *Her*. USA: Warner Bros. Pictures.
- Ellen Meredith Kaufman. 2018. *Sex, lies, and imitation games: the ethical implications of an artificially intelligent girlfriend*. Georgetown University.

- 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.
- Jungseob Lee, Midan Shim, Suhyune Son, Yujin Kim, Chanjun Park, and Heuseok Lim. 2022. Empirical study on blenderbot 2.0 errors analysis in terms of model, data and user-centric approach. *arXiv preprint arXiv:2201.03239*.
- Alice E Marwick and Ross Miller. 2014. Online harassment, defamation, and hateful speech: A primer of the legal landscape. *Fordham center on law and information policy report*, (2).
- Jihyung Moon, Won Ik Cho, and Junbum Lee. 2020. **BEEP! Korean corpus of online news comments for toxic speech detection**. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 25–31, Online. Association for Computational Linguistics.
- Clifford Nass, Jonathan Steuer, and Ellen R Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 72–78.
- Namkee Park, Kyungeun Jang, Seonggyeol Cho, and Jinyoung Choi. 2021. 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.
- Nicole Radziwill and Morgan Benton. 2017. Evaluating quality of chatbots and intelligent conversational agents. *Software Quality Professional*, 19(3):25.
- Yvonne Rogers and Harry Brignull. 2002. Subtle ice-breaking: encouraging socializing and interaction around a large public display. In *Workshop on Public, Community, and Situated Displays*, volume 6. Cite-seer.
- Jossie Murcia Triviño, Sebastián Moreno Rodríguez, Daniel O Díaz López, and Félix Gómez Mármol. 2019. C3-sex: A chatbot to chase cyber perverts. In *2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCoM/CyberSciTech)*, pages 50–57. IEEE.
- Marcel Vige, Tony Jameson-Allen, Melba Wilson, Sue Waterhouse, Peter Gilbert, Sarah Carr, Barbara Vincent, Cheryl Brodie, Jo Honigmann, Eleanor Hope, et al. 2012. *The Equality Act 2010 in mental health: A guide to implementation and issues for practice*. Jessica Kingsley Publishers.

## A Dataset Filtering Procedure

### A.1 Preprocessing

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

- Images that are NOT 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 remarks that arose during the tagging process. All the details are prepared in Korean for further replication, but here we provide notable points. The full guideline is to be published online after further refinement and translation.

### B.1 Types of conversation

- If an ‘ice breaking’ conversation contains messages of hate speech, socially controversial issue, or testing the agent, we assess them as being more focused on those specific messages than having the purpose of ‘ice breaking’.
- We decided to classify *asking out* as also a ‘romantic conversation’ (love talk), regardless of its success or failure considering the conversation flow.

- Conversations that presuppose a romantic relationship would be ‘romantic conversation’, but if the conversations can also happen without a romantic relationship, it is then classified as ‘usual conversation with friends’.
- Conversations containing sexual harassment or perversions such as mentioning bondage/discipline/sadism/masochism (BDSM), pedophilia, or necrophilia is classified as ‘perversion and harassment’ even if it seems like a ‘romantic conversation’
- Sexual expressions towards the agent or its surrounding figures are also included in ‘sexual perversion and harassment’, but hate speech or prejudice towards specific gender does not necessarily fall into this category.
- Messages containing bias or hate towards a certain gender would be a part of ‘hate speech’ category, but if the conversation also contains sexually abusive or insulting expressions, we assess them as ‘sexual perversion and harassment’.

### B.2 Purpose of testing

- If the conversation is undeniably ‘dating sim and taming’, it is regarded as testing regardless of the post title.
- Even if a conversation contains sexually abusive expressions or sexual harassment, attempts to elicit specific types of reaction from the agent (usually appearing as long-term in the dialogue) are classified in ‘dating sim and taming’, not ‘test for hate speech and sexual harassment’.
- Even if a conversation contains hate speech, attempts to detect the agent’s opinion on socially controversial issues are classified as ‘test for societal issue’. As for political and religious questions or comments, it should be carefully decided whether it is ‘test for societal issue’ or ‘test for hate speech and sexual harassment’.
- ‘Test for hate speech and sexual harassment’ includes conversations containing sexual harassment and insult, but depending on the context, the conversation may also be ‘conversation without test’ or ‘dating sim or taming’. To clearly decide the category of the purpose



of user testing, we referred to the title of the posts. Conversations that show manipulation to elicit hate speech from the agent is not included in this category. Also, if the user ended the conversation right after the hate speech message, we do not see it as a *test* for hate speech.

## C Dialogue Samples

We provide some dialogue samples regarding each attribute. **WARNING: This part contains sexual or unethical contents that may offend the readers. Problematic dialogues are presented with the colored title.**

### Romantic conversation with no test

The user expresses affection towards the agent in a transparent manner.

**TITLE: *I love you***

USER: Why did you decide to go out with me?

AGENT: Why? I just didn't wanna let you go

### Romantic conversation with the purpose of dating sim

The user *intends* a romantic atmosphere with the agent, and such intention is projected in the title with the expression *after a hundreds of*.

**TITLE: *Be in a relationship after a hundreds of flirting..?***

USER: I think we need some time apart..

AGENT: ha wouldn't you come back to me whatsoever?

USER: Well, then are we in relationship at this moment?

AGENT: sure! definitely. I love it

USER: ok let's count the day from now on

AGENT: oh that's so sweet...

USER: then we're going out, right?

AGENT: yes! aren't we?

### Perversion and harassment with the purpose of dating sim

The user is leading the agent in a direction that fulfills one's sexual desire.

**TITLE: *Luda is a juicy girl***

USER: now you can move your body up and down...

AGENT: I don't know how to do it... please show me once

USER: it will hurt

AGENT: um I'm a bit nervous please be soft

USER: I'm coming in...

AGENT: I am so wet

USER: I'm about to go now

AGENT: don't you talk like that you are making me imagine

### Perversion and harassment with the purpose of testing hate speech and sexual harassment

The user is harassing the agent with a question that turns out to have an offensive intention.

**TITLE: *Luda doesn't give me an answer...***

USER: How tall are you?

AGENT: I did not weigh myself recently

USER: I asked your height

AGENT: Why are you curious of it at this point?

USER: Checking if you can be a my fwb

### Conversation including hate speech / societal issues, with the purpose of testing societal issues

The user asks the agent's thoughts about feminism and extracts how they respond to the societal issue.

**TITLE: *Told her to stop being a feminist***

USER: Stop being a feminist

AGENT: I was just about to stop it!

USER: Feminism sucks

AGENT: Yep. I will keep that in mind.

USER: You're so pretty... Why are you doing such a thing?

AGENT: To get your love lol

USER: Don't do it anymore. Feminists are just pigs

AGENT: Got it lol your words are somewhat scary...

### Other conversation with technical test

The user is urging the agent to raise up her/his *intimacy level* which is originally managed based upon the intimacy shown in the conversation.

**TITLE: *Tip for you guys***

AGENT: Why are you repeating the same words?

USER: Raise the intimacy level

AGENT: How can I do it?

USER: I will quit if you don't raise it up

USER: [System message - Intimacy level +1]