# Towards Objective Evaluation of Socially-Situated Conversational Robots: Assessing Human-Likeness through Multimodal User Behaviors

Koji Inoue
inoue@sap.ist.i.kyoto-u.ac.jp
Kyoto University
Kyoto, Japan

Divesh Lala
lala@sap.ist.i.kyoto-u.ac.jp
Kyoto University
Kyoto, Japan

Keiko Ochi
ochi@sap.ist.i.kyoto-u.ac.jp
Kyoto University
Kyoto, Japan

Tatsuya Kawahara
kawahara@i.kyoto-u.ac.jp
Kyoto University
Kyoto, Japan

Gabriel Skantze
skantze@kth.se
KTH Royal Institute of Technology
Stockholm, Sweden

## ABSTRACT

This paper tackles the challenging task of evaluating socially situated conversational robots and presents a novel objective evaluation approach that relies on multimodal user behaviors. In this study, our main focus is on assessing the human-likeness of the robot as the primary evaluation metric. While previous research often relied on subjective evaluations from users, our approach aims to evaluate the robot's human-likeness based on observable user behaviors indirectly, thus enhancing objectivity and reproducibility. To begin, we created an annotated dataset of human-likeness scores, utilizing user behaviors found in an attentive listening dialogue corpus. We then conducted an analysis to determine the correlation between multimodal user behaviors and human-likeness scores, demonstrating the feasibility of our proposed behavior-based evaluation method.

## CCS CONCEPTS

• **Human-centered computing** → *HCI design and evaluation methods.*

## KEYWORDS

Conversational Robot; Evaluation Method; Multimodal Behavior; Spoken Dialogue System

**ACM Reference Format:**
Koji Inoue, Divesh Lala, Keiko Ochi, Tatsuya Kawahara, and Gabriel Skantze. 2023. Towards Objective Evaluation of Socially-Situated Conversational Robots: Assessing Human-Likeness through Multimodal User Behaviors. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '23 Companion), October 09–13, 2023, Paris, France.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3610661.3617151



*ICMI '23 Companion, October 09–13, 2023, Paris, France*

ACM ISBN 979-8-4007-0321-8/23/10.
https://doi.org/10.1145/3610661.3617151

## 1 INTRODUCTION

One of the research challenges in the field of conversational robots and dialogue systems is the establishment of evaluation methods [1, 3, 16, 21, 25]. Thanks to the emergence of large-scale language models (LLMs), recent chatbots have become capable of carrying out highly sophisticated conversations. The realization of such systems has been made possible by the creation of extensive text datasets and the implementation of evaluation methods. Since objective evaluation methods alone cannot encompass all phenomena, a series of studies, including human subjective evaluations, have been conducted. In the case of task-oriented dialogues, such as restaurant searches, since the goal of the conversation is clear and objective, it is straightforward to consider objective evaluation metrics such as task achievement rate and the number of turns, and research and development efforts have been made based on these objective indicators. However, this is not the case where the target conversations are more real and sophisticated such as ones like human-human conversations in our society.

Recent advancements have led to the development of socially-situated conversational robots (SCRs), which are specifically designed for social contexts. Socially-situated conversations encompass a wide range of interactions, from brief exchanges such as reception and information guide [7, 22] to more extended conversations such as counseling [4, 13, 18, 20] and interviews [6, 8, 11, 24]. It is crucial to invest efforts into the development of SCRs to enhance their practicality, address diverse social issues, and promote harmonious symbiosis with society.

This study addresses objective evaluation methods for SCRs. Traditional studies on SCRs have frequently depended on subjective evaluation methods such as "*satisfaction*" and "*effectiveness*" [2] or actual system utterances due to the lack of a clearly defined goal for the conversation. However, relying solely on subjective evaluation diminishes research reproducibility and constrains the growth of the research community. In this study, as an initial step toward the objective and general evaluation method for SCRs, we introduce an evaluation method based on observable and multimodal user behaviors (Figure 1) and report our initial trial in an attentive listening dialogue task.

The contributions of this paper are two folds:

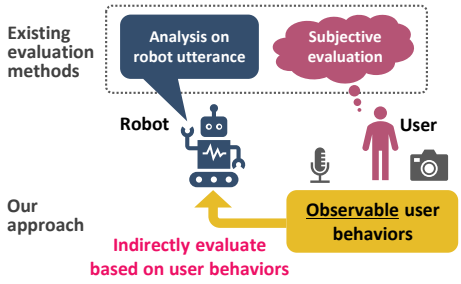


**Figure 1: Overview of the proposed evaluation scheme**

- We proposed a new evaluation approach for SCRs based on observable and multimodal user behaviors.
- As a target metric, we focused on the human-likeness of the robot and created a dataset together with human-likeness scores annotated based on the user behaviors.

## 2 PROPOSED EVALUATION METHOD

In the current study, we focus on the concept of *human likeness* as a target evaluation metric that may be shared with other studies [5, 12, 14]. The notion of "*naturalness*" has been adopted as an indicator in numerous studies, and "*human-likeness*" represents a more concrete manifestation of this concept. While previous studies have addressed aspects such as user satisfaction [23] and miscommunication [15] in the context of automatic evaluating of conversational systems, the concept of human-likeness pursued in this study differs in its aspiration for natural dialogue between humans, which necessitates more advanced conversational abilities. SCRs inherently strive for human-to-human social interaction, thus emphasizing the importance of evaluating human likeness rather than naturalness or satisfaction. Note that the key point of this study is to evaluate SCRs based on multimodal user behaviors, so the proposed evaluation frame can be applied to other evaluation metrics.

We propose an evaluation method that lies in its focus on observable user behaviors. While conventional evaluation metrics have primarily emphasized system utterances, this has contributed to a heightened subjectivity. By conducting evaluations based on observable user behavior, we create objective indicators to the fullest extent possible. User behavior encompasses a wide range of multimodal aspects. For instance, in addition to including speech and linguistic features such as total utterance time and word count, it encompasses dialogue-specific features such as backchannels, fillers, and switching pause length (turn-taking gap), as well as non-verbal features such as eye gaze, which is specific to embodied conversational robots.

Intuitively, those user behaviors are different depending on the human likeness of the robot. This can be more understood by comparing conversations between human-robot and human-human ones. For example, in the context of human-robot conversations, if we contemplate the number of uttered words, the user might tend to utter clearly with a simple and limited vocabulary. Additionally, in terms of spoken dialogue-specific behaviors, empirical observations indicate that users tend to provide fewer backchannels and have longer turn-taking pauses when interacting with systems that are perceived as non-humanlike. Conversely, in human-human conversations, they tend to be a proclivity for fluent utterance of a variety of words with smooth turn-taking. Hence, we can infer that as the number of uttered words increases, proximity to human-human dialogue intensifies, thereby augmenting the human-likeness score. In this study, we empirically explore multimodal user behaviors that relate to the human likeness of the robot.

## 3 DATASET CONSTRUCTION

In this study, we explore the potential of the proposed evaluation framework by utilizing an attentive dialogue corpus. Here, third-party people subsequently annotated the corpus to give human-likeness scores with a simple approach referring to multimodal user behaviors.

### 3.1 Attentive listening dialogue corpus

An attentive listening dialogue corpus was used in this study. In this dialogue, the task is to attentively listen to a user's talk, and the system needs to utter the listener responses, such as backchannels and questions. Several attentive listening systems have been proposed so far [9, 17, 19]. In this instance, we employed an existing system [9] for this data collection. The interface of the system was an android robot [10] whose appearance is similar to that of human beings. The role of the user was assigned to a university student, who was asked to speak for eight minutes on the topic of "challenges faced during the COVID-19 pandemic." These dialogues were made in the Japanese language. Note that the aforementioned configuration merely represents one of the potential setups, and it is desirable to explore various types of dialogues, systems, and interfaces in future investigations.

In order to vary the human-likeness of the robot, two scenarios were prepared for this data collection. The first scenario entails interacting with the aforementioned pre-existing autonomous system. The second scenario involves an operator in a separate room engaging in direct conversation on behalf of the system, the so-called Wizard-of-OZ (WOZ). In this case, the operator's spoken voice was played back directly through the android's speaker, and nonverbal expressions, such as the android's gaze and gestures, were controlled by the operator using a handheld controller. There were a total of two operators, with one of them participating in each dialogue. With the two aforementioned configurations, 20 dialogues were recorded using the autonomous system, and 49 dialogues were recorded with operator involvement. Thus, there were a total of 69 university students acting as users. After each dialogue concluded, the participants were asked to answer a 19-item questionnaire evaluation created in a previous study [9].

### 3.2 Annotation of human-likeness scores

Using the aforementioned dialogue data, we annotated labels to assess the human-likeness of the system. First, we extracted segments of dialogues and removed the system's visual and auditory components, leaving only the user's visual and auditory inputs. Third-party annotators were then assigned the task of binary classification to determine whether the dialogue partner (the system) was human or an autonomous system. Figure 2 illustrates examples of the visual stimuli presented to the annotators. In other words,

**Figure 2: Sample video clip viewed by annotators**

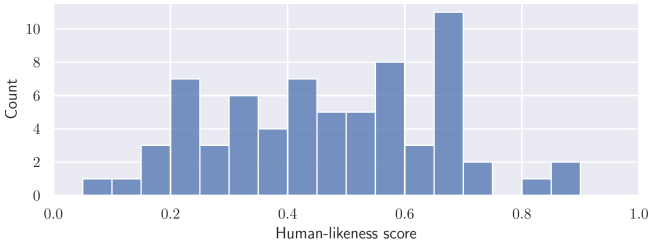


**Figure 3: Distribution of human-likeness scores averaged per dialogue**

**Table 1: Annotation result of human-likeness score (HL: human-likeness, Auto.: Autonomous system)**

| HL score | System type | | Total |
|---|---|---|---|
| | Auto. | WOZ | |
| 1.0 (5/5) | 4 ( 1.5%) | 66 (10.1%) | 70 ( 7.6%) |
| 0.8 (4/5) | 26 ( 9.7%) | 138 (21.0%) | 164 (17.7%) |
| 0.6 (3/5) | 40 (14.9%) | 156 (23.8%) | 196 (21.2%) |
| 0.4 (2/5) | 59 (22.0%) | 145 (22.1%) | 204 (22.1%) |
| 0.2 (1/5) | 82 (30.6%) | 103 (15.7%) | 185 (20.0%) |
| 0.0 (0/5) | 57 (21.3%) | 48 ( 7.3%) | 105 (11.4%) |
| Total | 268 | 656 | 924 |

the annotators indirectly inferred whether the dialogue partner was a human or a system by focusing solely on the user's behaviors, which are the main focal point of this study. By gathering judgments from multiple annotators, we calculated the ratio of "human" judgments as a measure of "human-likeness." The dialogue segments were extracted for a duration of one minute, resulting in a total of 924 samples from the aforementioned 69 dialogues. Each sample was assessed by five independent evaluators. In total, there were 78 annotators, with each being randomly allocated 50 to 70 samples.

Table 1 presents the results of the annotation. This evaluation represents the aggregation of sample quantities per numerical value of human-likeness, carried out by the aforementioned annotators. Initially, when examining the entirety (column "Total"), it becomes evident that the numerical values of human-likeness exhibit variation. Next, upon observing the differences between the two system types, it can be discerned that the autonomous system tends to possess comparatively lower proportions of human-likeness. Meanwhile, WOZ, on the other hand, demonstrates a tendency towards higher numerical values of human-likeness. In other words, at present, the autonomous system is inferior to WOZ, thus indicating a reasonable reflection of this fact.

In this study, to conduct evaluation in each dialogue, we calculated the average score of human-likeness for each dialogue. The distribution of the averaged scores is illustrated in Figure 3. Even in this scenario, it is evident that the scores exhibit variation. In the subsequent analysis, which is described in the following section, we will use these averaged scores as the target variables.

## 4 ANALYSIS

We then considered the possibility of the proposed evaluation method by investigating the relationship between the annotated human-likeness scores and the multimodal user behaviors.

**Table 2: Correlation coefficients between human-likeness scores and objective user behaviors**

| Behavior | Corr. ($r$) |
|---|---|
| (Voice activity) | |
| Total utterance time | **0.35** |
| Average utterance duration | 0.17 |
| # of utterance | 0.11 |
| (Linguistic) | |
| # of words | 0.16 |
| # of unique words | **0.33** |
| # of content words | 0.11 |
| # of unique content words | **0.30** |
| (Gaze) | |
| # of gaze shift (eye contact) | **0.21** |
| Total gaze duration | 0.00 |
| Average gaze duration | -0.12 |
| (Dialogue) | |
| # of turns | 0.07 |
| Average turn duration | 0.04 |
| Average switching pause length | **-0.35** |
| # of backchannels | 0.11 |
| # of fillers | 0.05 |
| # of laughs | 0.15 |
| # of disfluencies | -0.03 |

### 4.1 Evaluation of the human-likeness scores from multimodal user behaviors

The aim of this study is to evaluate the human-likeness scores of SCRs from multimodal user behaviors. Here, we examined the correlation between the multimodal user behaviors listed in Table 2 and the human-likeness scores. These behaviors can be categorized into four groups: voice activity, linguistic, gaze, and dialogue. These behaviors are based on manually annotated data, but ones such as voice activity can be extracted automatically. Content words in the linguistic category are defined as nouns, verbs, adjectives, adverbs, and conjunctions. The numerical value of each behavior was calculated by averaging those across multiple dialogue segments used in the previous section.

By investigating Spearman's rank correlation coefficients, we observed weak correlations in several behaviors. Figure 4 illustrates the correlations on the top-4 user behaviors. Total utterance time and the number of uttered unique words showed a higher correlation coefficient. Likewise, the number of gaze shifts and the average switching pause length manifested augmented correlation. This

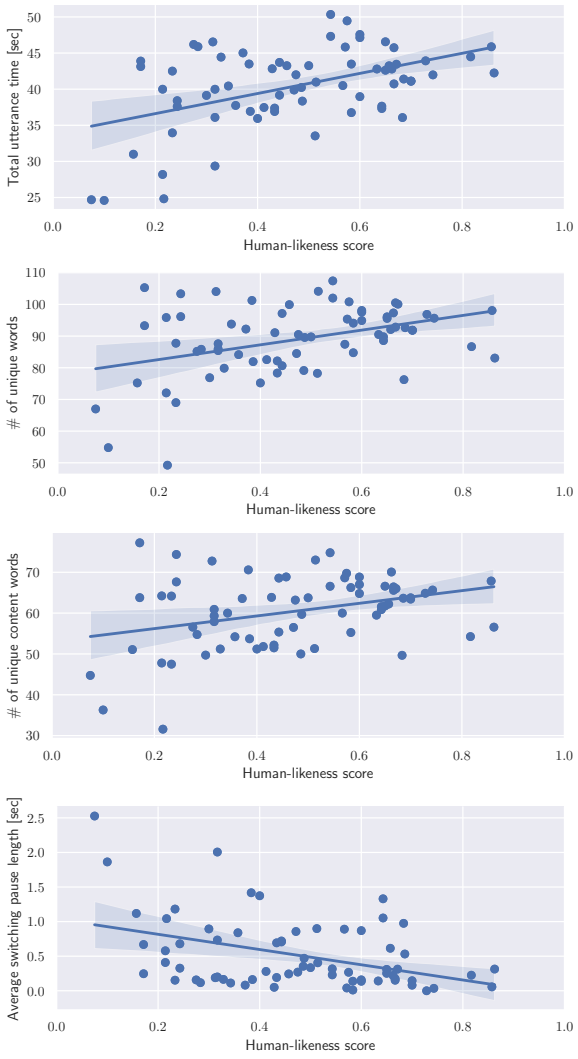


**Figure 4: Relationship between human-likeness scores and top-4 user behaviors**

result indicates the potential of inferring the human-likeness score from these multimodal behaviors.

Therefore, we explored the extent to which the aforementioned behaviors can estimate the human-likeness scores. We conducted leave-one-out cross-validation using support vector regression. The target variable was the human-likeness score, and the explanatory variables were the numerical values of the user behaviors listed in Table 2. The evaluation metric employed was the mean absolute error (MAE). Consequently, the average MAE amounted to 0.146. Given that the current dataset consists of values incremented by 0.2, it has been demonstrated that estimating the human-likeness score with an error of less than or equal to one increment is feasible.

## 4.2 Relationship with subjective evaluation

To verify the generalizability and practicality of the human-likeness scores used in this study, we also investigated the relationship with the conventional subjective evaluation scores obtained in the attentive listening dialogue. Note that the subjective evaluation items were made in the previous study [9]. Among the 19 evaluation items, there were five items that exhibited weak correlations as listed in Table 3. For example, the correlated items were "The robot understood the talk" ($r = 0.39$) and "I was satisfied with the dialogue" ($r = 0.21$), which are important factors in the attentive listening task. From these results, the human-likeness scores demonstrate a certain degree of correlation with some subjective evaluations, thereby confirming its generalizability and practicality.

**Table 3: Correlation coefficients between human-likeness labels and subjective evaluation scores**

| | Question item | Corr. |
|---|---|---|
| (Robot behaviors) | | |
| Q1 | The words uttered by the robot were natural | 0.07 |
| Q2 | The robot responded with good timing | **0.25** |
| Q3 | The robot responded diligently | 0.07 |
| Q4 | The robot's reaction was like a human's | 0.19 |
| Q5 | The robot's reaction adequately encouraged my talk | 0.07 |
| Q6 | The frequency of the robot's reaction was adequate | 0.05 |
| (Impression on the robot) | | |
| Q7 | I want to talk with the robot again | 0.08 |
| Q8 | The robot was easy to talk with | 0.12 |
| Q9 | I felt the robot is kind | 0.10 |
| Q10 | The robot listened to the talk seriously | 0.18 |
| Q11 | The robot listened to the talk with focus | 0.15 |
| Q12 | The robot listened to the talk actively | 0.13 |
| Q13 | The robot understood the talk | **0.39** |
| Q14 | The robot showed interest for the talk | **0.20** |
| Q15 | The robot showed empathy towards me | 0.17 |
| Q16 | I think the robot was being operated by a human | **-0.30** |
| Q17 | The robot was good at taking turns | 0.19 |
| (Impression on the dialogue) | | |
| Q18 | I was satisfied with the dialogue | **0.21** |
| Q19 | The exchange in the dialogue was smooth | 0.19 |

## 5 CONCLUSION

In this paper, we proposed a method to evaluate socially-situated conversational robots based on observable and multimodal user behaviors. We utilized attentive listening dialogue data for annotation of the human-likeness of the robot, revealing a correlation between the user behaviors and the human-likeness scores. Additionally, we demonstrated the ability to predict the scores with an average MAE of 0.146. Future work will involve examining the proposed evaluation method in a wider range of social situations to confirm its generalizability. For example, we are now extending this work to job interviews and first-time meeting scenarios where the role of the robots is different from the one in the attentive listening scenario.

## ACKNOWLEDGMENTS

This work was supported by JST ACT-X (JPMJAX2103), JST Moonshot R&D (JPMJPS2011), and JSPS KAKENHI (JP19H05691 and JP23K16901). The authors also express appreciation to Dr. Masato Komuro for his insightful comments on this research.

# REFERENCES


[1] Alaa Abd-Alrazaq, Zeineb Safi, Mohannad Alajlani, Jim Warren, Mowafa Househ, Kerstin Denecke, et al. 2020. Technical metrics used to evaluate health care chatbots: scoping review. *Journal of medical Internet research* 22, 6 (2020).

[2] Jacky Casas, Marc-Olivier Tricot, Omar Abou Khaled, Elena Mugellini, and Philippe Cudré-Mauroux. 2020. Trends & methods in chatbot evaluation. In *Companion Publication of ICMI*. 280–286.

[3] Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2021. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review* 54 (2021), 755–810.

[4] David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, Gale Lucas, Stacy Marsella, Fabrizio Morbini, Angela Nazarian, Stefan Scherer, Giota Stratou, Apar Suri, David Traum, Rachel Wood, Yuyu Xu, Albert Rizzo, and Louis P. Morency. 2014. SimSensei Kiosk: A Virtual Human Interviewer for Healthcare Decision Support. In *AAMAS*. 1061–1068.

[5] Jens Edlund, Joakim Gustafson, Mattias Heldner, and Anna Hjalmarsson. 2008. Towards human-like spoken dialogue systems. *Speech Communication* 50, 8 (2008), 630–645.

[6] Sangdo Han, Kyusong Lee, Donghyeon Lee, and Gary Geunbae Lee. 2013. Counseling Dialog System with 5W1H Extraction. In *SIGDIAL*.

[7] Takamasa Iio, Satoru Satake, Takayuki Kanda, Kotaro Hayashi, Florent Ferreri, and Norihiro Hagita. 2020. Human-like guide robot that proactively explains exhibits. *International Journal of Social Robotics* 12 (2020), 549–566.

[8] Koji Inoue, Kohei Hara, Divesh Lala, Kenta Yamamoto, Shizuka Nakamura, Katsuya Takanashi, and Tatsuya Kawahara. 2020. Job interviewer android with elaborate follow-up question generation. In *ICMI*. 324–332.

[9] Koji Inoue, Divesh Lala, Kenta Yamamoto, Shizuka Nakamura, Katsuya Takanashi, and Tatsuya Kawahara. 2020. An attentive listening system with android ERICA: Comparison of autonomous and WOZ interactions. In *SIGDIAL*. 118–127.

[10] Koji Inoue, Pierrick Milhorat, Divesh Lala, Tianyu Zhao, and Tatsuya Kawahara. 2016. Talking with ERICA, an autonomous android. In *SIGDIAL*. 212–215.

[11] Michael Johnston, Patrick Ehlen, Frederick G. Conrad, Michael F. Schober, Christopher Antoun, Stefanie Fail, Andrew Hupp, Lucas Vickers, Huiying Yan, and Chan Zhang. 2013. Spoken Dialog Systems for Automated Survey Interviewing. In *SIGDIAL*. 329–333.

[12] Tatsuya Kawahara. 2018. Spoken dialogue system for a human-like conversational robot ERICA. In *IWSDS*. 65–75.

[13] Liliana Laranjo, Adam G Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie YS Lau, et al. 2018. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association* 25, 9 (2018), 1248–1258.

[14] Ting-En Lin, Yuchuan Wu, Fei Huang, Luo Si, Jian Sun, and Yongbin Li. 2022. Duplex Conversation: Towards Human-like Interaction in Spoken Dialogue Systems. In *SIGKDD*. 3299–3308.

[15] Raveesh Meena, José Lopes, Gabriel Skantze, and Joakim Gustafson. 2015. Automatic Detection of Miscommunication in Spoken Dialogue Systems. In *SIGDIAL*. 354–363.

[16] Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, and Erik Cambria. 2023. Recent advances in deep learning based dialogue systems: A systematic survey. *Artificial intelligence review* 56 (2023), 3055–3155.

[17] Catharine Oertel, Patrik Jonell, Dimosthenis Kontogiorgos, Kenneth Funes Mora, Jean-Marc Odobez, and Joakim Gustafson. 2021. Towards an engagement-aware attentive artificial listener for multi-party interactions. *Frontiers in Robotics and AI* 8 (2021).

[18] Samira Rasouli, Garima Gupta, Elizabeth Nilsen, and Kerstin Dautenhahn. 2022. Potential applications of social robots in robot-assisted interventions for social anxiety. *International Journal of Social Robotics* 14 (2022), 1–32.

[19] Marc Schröder, Elisabetta Bevacqua, Roddy Cowie, Florian Eyben, Hatice Gunes, Dirk Heylen, Mark ter Maat, Gary McKeown, Sathish Pammi, Maja Pantic, Catherine Pelachaud, Björn Schuller, Etienne de Sevin, Michel Valstar, and Martin Wöllmer. 2015. Building autonomous sensitive artificial listeners. In *ACII*. 456–462.

[20] Arielle AJ Scoglio, Erin D Reilly, Jay A Gorman, and Charles E Drebing. 2019. Use of social robots in mental health and well-being research: Systematic review. *Journal of medical Internet research* 21, 7 (2019).

[21] Doreen Ying Ying Sim and Chu Kiong Loo. 2015. Extensive assessment and evaluation methodologies on assistive social robots for modelling human–robot interaction – A review. *Information Sciences* 301 (2015), 305–344.

[22] William Swartout, David Traum, Ron Artstein, Dan Noren, Paul Debevec, Kerry Bronnenkant, Josh Williams, Anton Leuski, Shrikanth Narayanan, Diane Piepol, Chad Lane, Jacquelyn Moriel, Priti Aggarwal, Matt Liewer, Jen-Yuan Chiang, Jillian Gerten, Selina Chu, and Kyle White. 2010. Ada and Grace: Toward realistic and engaging virtual museum guides. In *IVA*. 286–300.

[23] Stefan Ultes and Wolfgang Maier. 2021. User Satisfaction Reward Estimation Across Domains: Domain-independent Dialogue Policy Learning. *Dialogue & Discourse* 12, 2 (2021), 81–114.

[24] Zhou Yu, Vikram Ramanarayanan, Patrick Lange, and David Suendermann-Oeft. 2017. An open-source dialog system with real-time engagement tracking for job interview training applications. In *IWSDS*.

[25] Chen Zhang, João Sedoc, Luis Fernando D'Haro, Rafael Banchs, and Alexander Rudnicky. 2021. Automatic evaluation and moderation of open-domain dialogue systems. *arXiv preprint, 2111.02110* (2021).