# Exploring User Engagement Through an Interaction Lens:

## What Textual Cues Can Tell Us about Human-Chatbot Interactions


Linwei He
Department of Communication and Cognition, Tilburg School of Humanities and Digital Sciences, Tilburg University, the Netherlands
l.he_1@tilburguniversity.edu

Anouck Braggaar
Department of Communication and Cognition, Tilburg School of Humanities and Digital Sciences, Tilburg University, the Netherlands
a.r.y.braggaar@tilburguniversity.edu

Erkan Basar
Behavioral Science Institute, Radboud University, Nijmegen, the Netherlands
erkan.basar@ru.nl

Emiel Krahmer
Department of Communication and Cognition, Tilburg School of Humanities and Digital Sciences, Tilburg University, the Netherlands
e.j.krahmer@tilburguniversity.edu

Marjolijn Antheunis
Department of Communication and Cognition, Tilburg School of Humanities and Digital Sciences, Tilburg University, the Netherlands
m.l.antheunis@tilburguniversity.edu

Reinout Wiers
Addiction Development and Psychopathology (ADAPT)-lab, Department of Psychology, University of Amsterdam, the Netherlands
r.w.wiers@gmail.com



## ABSTRACT

Monitoring and maintaining user engagement in human-chatbot interactions is challenging. Researchers often use cues observed in the interactions as indicators to infer engagement. However, evaluation of these cues is lacking. In this study, we collected an inventory of potential textual engagements cues from the literature, including linguistic features, utterance features, and interaction features. These cues were subsequently used to annotate a dataset of 291 user-chatbot interactions, and we examined which of these cues predicted self-reported user engagement. Our results show that engagement can indeed be recognized at the level of individual utterances. Notably, words indicating cognitive thinking processes and motivational utterances were strong indicators of engagement. An overall negative tone could also predict engagement, highlighting the importance of nuanced interpretation and contextual awareness of user utterances. Our findings demonstrated initial feasibility of recognizing utterance-level cues and using them to infer user engagement, although further validation is needed across different content-domains.



## CCS CONCEPTS

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

## KEYWORDS

User engagement, Human-chatbot interaction, Conversational agents

**ACM Reference Format:**
Linwei He, Anouck Braggaar, Erkan Basar, Emiel Krahmer, Marjolijn Antheunis, and Reinout Wiers. 2024. Exploring User Engagement Through an Interaction Lens:: What Textual Cues Can Tell Us about Human-Chatbot Interactions. In *ACM Conversational User Interfaces 2024 (CUI '24), July 08–10, 2024, Luxembourg, Luxembourg*. ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3640794.3665536





*CUI '24, July 08–10, 2024, Luxembourg, Luxembourg*



ACM ISBN 979-8-4007-0511-3/24/07
https://doi.org/10.1145/3640794.3665536


## 1 INTRODUCTION

With the rapid development of artificial intelligence (AI), chatbots - text-based conversational agents that simulate interactions with human users through natural language [25] - have become popular in view of their potential to provide accessible, autonomous, and personalized interactions. They have been applied in diverse domains such as healthcare, social companionship, and education. Recent research highlights the initial effectiveness and user acceptance of chatbots in motivating health behaviors [27], facilitating health education [24], and supporting mental health [1]. However, a notable limitation is that users tend to disengage with the chatbots over time [2], limiting their long-term effectiveness. This disengagement is particularly significant in contexts that involve multiple interactions, as is often the case in health intervention, where sustained engagement with the chatbot is a prerequisite for achieving behavior change outcomes [47]. In light of these observations, understanding and facilitating user engagement has become a major theme of research in chatbots and human-computer interaction (HCI) in general.

Engagement is a versatile concept with various interpretations emerging from the different contexts in which it has been studied. Engagement, as perceived and experienced, focuses on subjective feelings such as interest, excitement, and arousal [34, 46]. This is often measured through retrospective self-report methods such as questionnaires and interviews [21]. The behavioral aspect of engagement tends to emphasize active action and participation [21, 55], providing observable and objective measures such as visual and textual cues. Moreover, engagement is frequently characterized as a dynamic state, implying that it may change over time within an interaction [44]. It is, therefore, essential to monitor engagement throughout the interaction, enabling the system to maintain the engaged state and adapt in the event of disengagement. However, this poses a great challenge due to the retrospective nature of self-report measures and potentially interrupted interaction experience

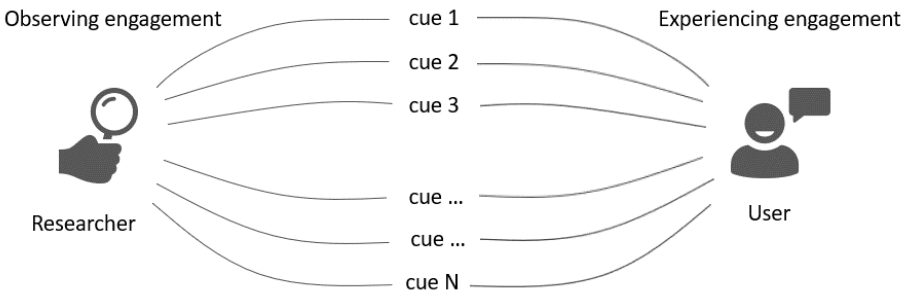


**Figure 1: A dual perspective lens framework of user engagement.**

(e.g., ecological momentary assessment where users repeatedly self-report during or immediately after an experience). Therefore, behavior observations are often used as proxies of user engagement. Researchers look for cues in the interaction and attempt to infer user engagement through them. These cues, such as length of utterances and use of certain words, serve as a lens through which user engagement is approximated. Figure 1 presents a lens framework, inspired by Brunswick's Lens model [12], illustrating and distinguishing between researchers' perceptions of engagement and users' experience of engagement. The original Lens model was developed to understand human judgment and decision-making processes. It posits that individuals make judgments through a lens, where they interpret cues (i.e., pieces of information or data that individuals gather and use to make judgments) based on their own personal knowledge, perception, and beliefs. This framework provides a useful analytic approach in understanding the interplay between the observation and the display of user engagement in human-chatbot interactions.

While using the cues as an engagement measure is efficient and reduces user burden, potential discrepancies may exist between researchers' observations and users' actual experience of engagement. As highlighted by Troy Frensley and colleagues [58], the observer's perception of engagement may differ from that of the user. A classic example is time spent on the content, which can be observed as a cue for engagement while the user might be experiencing confusion or frustration instead [9]. In a more recent comparison between objective and subjective measures of engagement, Silvervarg and colleagues [52] found that the behavioral cues (such as asking questions to the agent) did not adequately represent users' subjective feelings of engagement, emphasizing the need for further validation of the cues. This lens perspective has been actively used in human-agent interactions, especially within embodied agents, where researchers have established reliable cues such as gaze and head movement. However, to our knowledge, there is a lack of systematic evaluation of textual cues in the context of human-chatbot interaction. It remains unclear which textual cues can be used as proxies to infer user engagement. Moreover, there is a need for more research on testing the values of these cues. As suggested by [21], the ecological validity of these cues remains to be explored, such that whether the cues adequately correspond to user engagement and whether they can be generalized to different domains.

To address these methodological gaps, this study aims to answer the following research questions:

RQ1: What are the textual cues in the current research field of human-chatbot interaction?

RQ2: What is the relationship between these cues and user engagement? In other words, do these cues correspond to user engagement?

We present a three-fold contribution to the understanding and evaluation of user engagement in text-based chatbot interactions. To answer RQ1, we compiled an inventory of textual engagement cues that have been used in prior research, specifically at the utterance level. This inventory provides an overview of methods for collecting and analyzing interaction data that signals engagement. To answer RQ2, we applied these textual cues on the annotation of a dataset comprising of 291 human-chatbot interactions, demonstrating the feasibility of recognizing engagement from the utterances. Taking a step further, we then explored the relationship between the cues and user engagement measured through self-report questionnaire, providing initial insights into whether these textual cues actually reflect users' subjective feelings of engagement. This exploration serves as a useful step toward a more comprehensive understanding and evaluation of user engagement in human-chatbot interactions.

## 2 RELATED WORK

Given that the evaluation of chatbots and dialogue systems in general is a long-lasting challenge, we aim to first identify approaches that have been used to understand user engagement. We organize the literature into three sections: (1) the application and evaluation of chatbots; (2) the various methods to measure engagement, including objective cues and users' experience; (3) the relationship between the two – how they have been combined and how they correspond with each other.

### 2.1 The application and evaluation of chatbots

Text-based chatbots are conversational agents that interact with users through natural language. The first known chatbot ELIZA

was developed in 1966 as a psychotherapist [51]. It used simple keyword matching technique and demonstrated that text-based communication between a human and a machine was possible [10]. In the recent decade, with the rapid advancement in AI technology, the application of chatbots has become exponentially pervasive. They have been applied in various domains, ranging from virtual customer service to social companionship [23]. In light of the application domains and purposes, Følstad and colleagues [23] developed a typology of chatbots, classifying them into long-term vs. short-term interaction and user-driven vs. chatbot-driven interaction. Customer support chatbots is a prominent example of short-term user-driven type, which often operate within a rule-based, domain-specific framework, relying on pre-defined dialogue flows to help users achieve certain goals. Recent advancements in Natural Language Processing (NLP) enables these chatbots to understand user input and extract relevant information, improving efficiency and accuracy [54]. On the other side of the spectrum, long-term chatbot-driven type can be found in domains such as healthcare and education. These chatbots prioritize creating a natural and engaging dialogue while modelling user profiles, with the goal of establishing and maintaining a long-term relationship in order to achieve behavior change, as an example.

With the growing popularity and wide usage domains of chatbots, their evaluation has become a major research theme. Researchers have focused on different aspects when evaluating chatbots, ranging from task performance to user satisfaction. Consequently, the evaluation methods also vary in accordance with these diverse aspects. In a recent review, Casas and colleagues [14] synthesized past research on chatbot evaluation and summarized three perspectives on the evaluation. First, *effectiveness or performance evaluation*, usually from a more technical point of view, concerns the performance of the underlying computational techniques. Example metrics includes accuracy and coherence and is often evaluated through combinations of human evaluation and automatic approaches such as F-scores [14]. The second perspective, *efficiency evaluation*, corresponds to the explicit goal of the chatbot. For example, whether the chatbot provides adequate information to the user and answers user queries correctly. The third perspective involves *satisfaction evaluation*, concerning the end users' experience with the chatbot. Engagement is one of the major aspects within this evaluation. The review noted that efficiency is currently the mostly used criteria for evaluating chatbots and calls for more research into further develop large-scale satisfaction evaluation of chatbots. This call closely matches with other research [47] suggesting that users need to first feel engaged with the system before they engage in the long-term goal, be it education or behavioral change. In the next section, we map out how user engagement has been evaluated in the literature.

## 2.2 User engagement

User engagement is thus crucial to both the design and evaluation of human-chatbot interactions. Previous HCI research has described several theories and definitions of engagement. A prominent perspective stems from the Flow theory [18], which characterizes engagement as an immersive and enjoyable experience. Within this framework, researchers distinguish between the cognitive, emotional, and behavioral aspects of engagement [21]. Cognitive engagement involves conscious elements such as attention, awareness, and effort [34]. Engagement has also been viewed as an emotional or affective process, emphasizing the subjective nature of the experience. An engaged state is reflected by feelings of interest and achievement, while disengagement is signaled by negative emotions such as boredom and frustration [26]. Lastly, the behavioral dimension of engagement entails active involvement and action [5], allowing observable and quantifiable assessment. Moreover, engagement is often described as a variable state rather than a stable characteristic [44]. Depending on various factors (e.g., the interaction, the environment, the context, the user), there could be temporal changes in engagement throughout the interaction. For instance, Yu and colleagues [63] designed an agent capable of recognizing disengagement through users' speech and facial expressions. The agent can then adapt its responses to re-engage users using strategies such as referring to shared experience from earlier conversations, creating a sense of common ground. This highlights the fluid nature of engagement, as individuals may experience moments of disengagement but can be successfully re-engaged with adaptations. The ability to recognize the dynamic states and facilitate smooth transitions between them is crucial in maintaining user engagement and encouraging continuous usage.

In light of the diverse interpretations of engagement, its measurement involves a variety of approaches. Cognitive and emotional engagement are often assessed through subjective methods, such as questionnaires, interviews, and other forms of self-report assessment [21]. They capture a reliable ground truth metric for user engagement, with the advantages of being easy to administer and reproducible. The User Engagement Scale (UES) is one of the most frequently used measures [43]. It operationalizes engagement as users' subjective feelings of attention, involvement, and perceived usability. While it may well reflect the theoretical components of engagement, its retrospective nature introduces a delay in measurement and cannot capture the immediate experiences. Experience Sampling Methods (ESM), another subjective-oriented approach, can monitor changes in engagement through repeated self-reporting [50]. It addresses the temporal and dynamic nature of engagement, allowing users to reflect on their momentary experiences and capturing a more comprehensive picture of the engagement process. Yet, ESM's limitations include repeated disruptions and increased user burden, potentially leading to disengagement [60].

In contrast, objective measures infer engagement without direct user contact, reducing burden and minimizing disruption [21]. These range from simple measures such as behavior logging to more elaborate human annotation. Example proxies include conversation length (e.g., number of utterances) and conversation depth (e.g., users' self-disclosure) [61]. Human annotation is a well-established method in evaluating dialogues systems, especially with embodied agents. For instance, mutual eye gaze is found to be a consistent indicator of engagement [41], while shift of visual focus is usually a sign of disengagement [49]. However, there is comparatively less annotation on text-based human-machine interactions, and there is currently a lack of established and validated measures in this domain. While conversation-level proxies like interaction duration

and number of utterances are frequently used to predict dropout (a measure of disengagement) [8], more detailed indicators are less explored. The advantage of observing utterance-level cues lies in that it responds to the momentary dimension of engagement, alerting the chatbot to adapt in the event of disengagement. However, unlike well-established questionnaires, there has not been comprehensive testing of these utterance-level indicators, calling for more systematic summarization and validation, and we aim to address this methodological gap in the current paper.

### 2.3 Infer engagement through a lens of textual cues

Due to the retrospective nature in self-report measures and the challenges in monitoring engagement in real time, interaction cues observed from user input provide a useful lens through which researchers infer whether users are engaged in the interaction. The simplest session-level cues include session length and number of turns [7]. While these cues are scalable and easy to measure, they have been criticized for their potentially shallow interpretation. For instance, a user spending a long time on a webpage might be frustrated instead of engaged [17, 21]. In extension to these, attempts have been made to infer engagement from more thin-sliced cues. Carlton and colleagues [13] analyzed user behaviors with narrative TV and found that proactively choosing topics (actions such as skipping content or seeking backwards) significantly predicted self-report engagement. In the domain of embodied agents, gaze and gestures are frequently used as proxies of engagement, and [33] found that engaged users, as measured through self-report, behaved more collaboratively and had more focused gaze. Similarly, users with more eye fixation found the agent more helpful [19]. However, unlike research in embodied agents where human annotation is widely used, there is currently no well-recognized standard for text-based chatbots. A few studies have explored utterance-level textual cues that are indicative of engagement, such as users' spontaneous questions and complaints about chatbot responses [39, 57]. However, these studies did not include subjective measures to cross-validate the value of these cues, and there may be a mismatch between the cues and actual user engagement, reflecting a lack of ecological validity. For instance, in an explicit comparison between subjective and objective measures, Silvervarg and colleagues [52] evaluated a conversational agent in learning environment and found that the behavioral cues (such as questions to the agent) did not adequately represent users' satisfaction and engagement.

In this study, we extend previous research and by compiling an inventory of textual cues that may be indicative of engagement and exploring the relationship between these cues and users' experiences of engagement.

## 3 METHODS

We followed a two-step approach for the investigation of user engagement. First, we carried out a rapid literature search to look for engagement cues that have been identified in the domain of text-based dialogue systems. Subsequently, these cues were annotated within a dataset consisting of 291 human-chatbot conversations, accompanied by self-report engagement scores measured via the User Engagement Scale (UES). We then used correlation and regression analysis to explore the relationship between the cues and UES scores.

### 3.1 An inventory of textual engagement cues

We developed two sets of search terms used for the literature search. The first set focused on chatbots and included synonyms such as "chatbot", "conversation* agent", and "dialogue system". The second set addressed user engagement and included keywords such as "engage*", "predict*", "recogniz*", "indicat*", "evaluat*", and "annotat*". A search of literature was then performed using these search terms, without constrains on publication time. Studies were included if they were published in English and if they: (1) had been peer-reviewed; (2) reported on the use of a text-based dialogue system; and (3) explicitly mentioned textual cues that are used to measure or annotate user engagement. The search was performed on Google Scholar and reference lists of relevant studies were also reviewed to identify additional relevant literature.

The search revealed that the majority of studies evaluating engagement were conducted with embodied agents, focusing on visual and speech annotations. There are considerably fewer studies that explicitly looked at text-based agents and engagement cues. Many studies focused on building prediction models of engagement using the cues, without including self-report as reference. Some papers took a more social-science perspective and compared people's use of the cues in comparison with human-human communication, touching upon broader concepts related to engagement, such as user satisfaction. Among these studies, interaction features such as session length and number of utterances are still the most frequently used cues [31, 57, 61, 64]. In terms of human annotations of more thin-sliced cues, Liang and colleagues [39] summarized a list of heuristic textual cues of user responses indicating disengagement. These include complaints about system responses (e.g., "you already told me that") and dislike the current topic (e.g., "I don't like this conversation"). Trinh and colleagues [57] built a prediction model for user engagement using predictors such as user asking questions to the chatbot. See Table 1 for an overview of the cues we compiled from the literature. We noticed little overlap in the cues used in different studies, which provides an opportunity to compare earlier findings in a different domain. Since the dataset used in this study was from a chatbot designed to discuss and motivate people to quit smoking, we also included cues specifically relevant to this domain, such as change talk (i.e., utterances indicating a positive attitude towards behavior change), sustain talk (i.e., utterances indicating a negative attitude towards behavior change), and self-reflection (i.e., utterances indicating active thinking on one's behaviors).

Additionally, some studies examined user experience of human-chatbot interactions by analyzing conversations through linguistic analysis. For example, [31] and [22] used the Linguistic Inquiry and Word Count program (LIWC) [56] to analyze people's use of language when communicating with a chatbot, comparing it with communication with a human. Their results suggest that people used more affect words, social and cognitiveprocess words with a chatbot, and experienced more positive emotion with the chatbot than with a human. LIWC enables automatic analysis of natural language data, providing insights into cognitive and psychological

**Table 1: Engagement cues and explanations.**

| Cue group | Engagement cue | Description | Examples |
|---|---|---|---|
| Utterance features | Complain bot repetition [39] | User complains about bot repeating itself | "you already told me that" |
| | Complain bot ignoring [39] | User complains about bot ignoring user | "You're not listening" |
| | Complain bot misunderstanding [39] | User complains about bot misunderstanding user | "That's not what I mean" |
| | Not understanding bot [39, 57] | User implies/seeks clarification | "what do you mean?" |
| | Express frustration [39] | User explicitly express frustration | "sigh", "ugh" |
| | Show low interest [39] | User shows low interest in the conversation (not in behavior change, in our context) | "just give me the code for completion" |
| | Express negative opinion about the conversation [39] | User expresses negative opinion about the conversation (not about behavior change, in our context) | "you ask too many dumb questions" |
| | Restatement [37, 38] | User restates their previous utterance | Could be repeating, rephrasing, adding or removing words |
| | User question [57] | User asks spontaneous questions to the chatbot | "what is the best method to quit smoking?" |
| | Acknowledgement [59] | User acknowledges the chatbot's suggestions, information, proposition | "Sure, let's talk about it" |
| | Shorthand [31] | an informal shortened-form language, including abbreviations and the omission of auxiliary verbs and pronouns | "Idk", "Wtv" |
| | Emoticon [31] | Emoticons are emotional icons used to represent a specific emotional state | ":/" ":D" |
| | Self-reflection [16, 42] | Reflective description (low-level), thoughts, (medium-level), feelings and evaluations (deep-level) one has about their own behaviors and experiences. | "I tried quitting 2 years ago, and I did feel lighter back then" |
| | Change talk [42] | Statements from the user that favors quitting. | "I'll start with throwing away all my cigarettes" |
| | Sustain talk [42] | Statements from the user that favors quitting. | "I don't smoke that much. I think I'm fine" |
| LIWC features [36] | Analytic | Metric of logical, formal thinking | -[a] |
| | Clout | Language of leadership, status | |
| | Authentic | Perceived honesty, genuineness | |
| | Positive tone | Degree of positive tone in general | |
| | Negative tone | Degree of negative tone in general | - |
| | Positive emotion | Words indicating positive emotions | Good, love, happy, hope |
| | Negative emotion | Words indicating negative emotions | Bad, hate, hurt, tired |
| | Cognitive process | Words indicating active cognitive process | But, think, because, remind |
| | Social process | Words indicating social behaviors, relations, and processes | Help, thank, tell, we, friend |
| Interaction features | Number of utterances | | |
| | Average utterance length | | |
| | Total interaction duration | | |

[a] These are overall metrics without specific word dictionaries.

processes through linguistic displays. Therefore, we also included major LIWC categories to explore the relationship between words and engagement [22, 36].

## 3.2 The dataset

The extracted engagement cues were then annotated in a conversational dataset comprising dialogues between a chatbot and 291 users. The dataset was from an earlier study in which participants interacted with a chatbot, Roby [28]. Roby was designed to chat with smoker users about their smoking behavior, aiming at motivating people to quit smoking. The chatbot operated on the Rocket.Chat web interface [66], which allows researchers to collect and store chat logs on a local server. The chatbot was scripted by the authors and was reviewed by a clinical psychology expert

to ensure appropriateness of its interactions. Users could interact with the chatbot using free-form text. Detailed descriptions of the content and technical infrastructure are published elsewhere [6, 28]. The chatbot featured two versions using different communication styles. One followed the motivational interviewing (MI) techniques, where the chatbot emphasized acceptance and empathy by asking open-ended questions and reflecting on user input. The other style incorporated therapeutic confrontation by providing more direct and confrontational feedback. The content of the dialogues and the length of the interaction was mostly comparable between the two versions. The study participants were university students aged 18 or older and current smokers. They were randomly assigned to one of the Roby versions and were instructed to initiate the interactions at their preferred time and place. After the interactions, participants completed a questionnaire evaluating their attitudes towards quitting smoking and their experiences with the chatbot. Self-report engagement were measured using the 5-point User Engagement Scale (UES) [43].

## 3.3 Analysis

*3.3.1 Annotation.* Manual annotation was applied to engagement cues extracted from the literature review. An annotation scheme was developed by the author team, specifying annotation segments and definitions of the cues. The annotation was performed per user utterance, and the occurrence count for each cue was recorded. For example, a user utterance like "You already told me that, but I don't want to quit" was counted as 1 for the cue "complain bot repetition" and 1 for the cue "sustain talk". First, in the iterative training phase, two annotators independently coded a number of randomly selected transcripts, and any disagreement were resolved through group discussions. In the third and last training session (15 transcripts, 5% of the total), the annotators reached a Krippendorff's Alpha of 0.74, indicating acceptable agreement [35]. The remaining transcripts were then divided between the two annotators.

LIWC variables were processed automatically via the LIWC program. For this purpose, the transcripts contained only user utterances, and the chatbot utterances were removed. Words in each transcript were matched with the ~6400 words contained in the LIWC dictionary [11], and the output number reflects the frequency of certain word categories (e.g., the frequency of words indicating cognitive process) or an aggregated rating of certain language variables (e.g., the rating of positive tone ranging from 0 to 100). See Table 1 for the LIWC variables included in our analysis and Table 3 for the output. We focused on these major variables since they provide insights into the emotional, cognitive, and structural components present in individual's language use, as suggested by previous studies in human-chatbot interaction [22].

*3.3.2 Analysis.* The goal of the analysis was to explore relationships between self-report engagement and textual engagement cues. As described above, these cues were organized into three groups: annotated variables, LIWC variables, and interaction variables. We followed a multi-step analysis plan to unveil the relationship between the cues and UES scores.

First, to discover the association between the cues and subjective engagement, correlation analyses were performed. This gives an initial overview, highlighting relevant cues and their potential relationship with UES, although not how they combined. To determine which correlation test to perform, we performed Shapiro-Wilk's test for normality. For normally distributed cues, we used Pearson's r correlation coefficients. For non-normally distributed cues, we calculated Spearman's rank-order coefficients.

The second step aimed to gain a more complete picture of whether latent relationships are actual indicators of engagement. We performed a hierarchical regression integrating cues that are associated with UES. We chose hierarchical regression because the three cue groups (i.e., the annotated variables, LIWC variables, and descriptive variables) are theoretically and practically different, and hierarchical regression allows the researcher to compare theory-informed models and understand the relative importance of the cues [32]. This comprehensive analysis, while controlling for demographics and domain-specific smoking-related variables, allowed us to explore the relative contributions of individual cues and groups, providing a more thorough overview of the relationship between textual cues and experiential engagement.

# 4 RESULTS

## 4.1 Sample

In total, 291 transcripts were annotated. Some participants did not finish the interaction and therefore did not complete the questionnaire afterwards, resulting in 202 participants with a self-report UES score. Overall, participants were moderately engaged with the chatbot ($M$=3.21, $SD$=0.76). The distribution of UES scores is shown in Figure 1. Since a large number of user utterances consisted of single or short phrases, such as responses to multiple choices questions, not all utterances contained cue occurrences. In total, we observed 5,572 cue occurrences across 10,554 utterances. Of the 5,572 cue occurrences, the majority featured positive responses to the chatbot, including *self-reflection* statement ($n$=722, 13.0%), *acknowledgement* ($n$=971, 17.4%), and *restatement* ($n$=656, 11.8%). Negative responses such as complaints occurred considerably less frequently. See Table 2,Table 3, and Table 4 for frequencies of all annotated cues.

Some of the cues (indicated by an * in Table 2) have a notably high frequencies of zero counts, indicating that they had very low occurrence. For instance, out of 291 interactions, only 4 contained *complain bot repetition*. Due to limited observations of these cues and therefore limited interpretability in statistic tests, we report them as descriptives and discuss possible implications of this in the Discussion.

## 4.2 Exploring relationships between cues and engagement

As a first step, we examined the specific cues that were associated with self-report UES scores. Correlation analyses were conducted between all the cues and UES score.[1] We found various correlations between the cues and USE, both positive and negative. Within the utterance feature group, cues correlated with UES included

[1]Please note that here we merely offer descriptive information on the strength of individual correlations as a first step towards the prediction of engagement in the next subsection. The reported, uncorrected p-values should therefore be interpreted with a grain of salt.

**Table 2: Descriptive results for engagement cues – Utterance features.**

| **Utterance features** | *n (%)* |
|---|---|
| Complain bot repetition* | 5 (0.09%) |
| Complain bot ignoring* | 10 (0.18%) |
| Complain bot misunderstanding* | 6 (0.11%) |
| Not understanding bot* | 42 (0.75%) |
| Express frustration* | 36 (0.65%) |
| Show low interest* | 38 (0.68%) |
| Express negative opinions about the conversation* | 4 (0.07%) |
| Restatement | 656 (11.8%) |
| User question* | 33 (0.59%) |
| Acknowledgement | 971 (17.4%) |
| Shorthand* | 22 (0.39%) |
| Emoticon * | 10 (0.18%) |
| Low-level self-reflection | 722 (13.0%) |
| Medium-level self-reflection | 349 (6.26%) |
| Deep-level self-reflection | 227 (4.07%) |
| Change talk | 578 (10.37%) |
| Sustain talk | 561 (10.7%) |

**Table 3: Descriptive results for engagement cues – Linguistic (LIWC) features.**

| **Linguistic (LIWC) features** | *M (SD)* |
|---|---|
| Analytic | 28.86 (19.62) |
| Clout | 1.82 (2.60) |
| Authentic | 61.72 (33.48) |
| Cognitive process | 11.86 (4.61) |
| Positive tone | 4.30 (2.12) |
| Negative tone | 1.49 (1.47) |
| Positive emotion | 0.65 (0.95) |
| Negative emotion | 0.92 (1.14) |
| Social process | 5.97 (3.96) |

**Table 4: Descriptive results for engagement cues – Interaction features.**

| **Interaction features** | *M (SD)* |
|---|---|
| Number of utterances | 36.62 (12.84) |
| Average utterance length (in words) | 2.43 (1.22) |
| Total interaction duration (in seconds) | 591.09 (571.40) |

*acknowledgement* ($r$ = 0.18, $p$ = .012), self-reflection, specifically *low-level self-reflection* ($r$ = 0.21, $p$ = .003) and *medium-level self-reflection* ($r$ = 0.26, $p$ <.001), *change talk* ($r$ = 0.31, $p$ < .001), and *sustain talk* ($r$ = -0.19, $p$ = .006). In addition, within the LIWC features, *authentic* ($r$ = 0.14, $p$ = .048), *cognitive process* ($r$ = 0.17, $p$ =.014), *negative tone* ($r$ = 0.15, $p$ = .034), and *positive emotion* words ($r$ = 0.18, $p$ = .009) were found to be positively associated with UES score. Lastly, regarding the interaction features, *number of utterances* ($r$ = 0.15, $p$ = .029), and *average utterance length* ($r$ = 0.18, $p$ = .010) correlated with UES score.

## 4.3 Exploratory prediction of engagement

The correlation analyses provide insights into the initial latent relationship between the cues and user engagement, answering our RQ2. To further explore whether these cues are predictive of engagement, we built hierarchical regression models to examine the overall combined effects of the cues. To account for differences in ranges of different measures, we scaled the variables prior to the regression. It should be noted that several cues were found to correlate with each other, leading to multicollinearity issue. One way to reduce multicollinearity, as suggested by O'brien [45], is to combine predictor variables that are conceptually similar and share high correlation into a single measure. Therefore, before

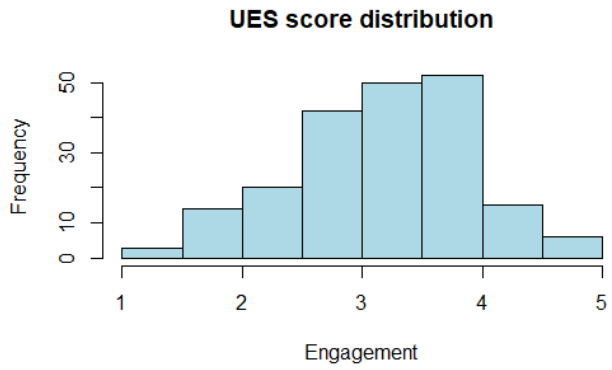


**Figure 2: Distribution of engagement scores.**

conducting the hierarchical regression, we aggregated some of the cues: Low- and medium-level self-reflection were combined into a single cue *self-reflection* by summing their counts. Change talk and sustain talk were aggregated into a single cue *motivational talk* by subtracting counts for sustain talk from that for change talk, capturing a relative motivation.

Next, these cues were integrated into a hierarchical regression to examine their combined effect and relative contribution to UES score. To account for individual differences in demographics and smoking behavior as well as the two chatbot versions, these variables (i.e., age, gender, years of smoking, daily cigarette consumption, nicotine dependence level, baseline motivation to quit, and chatbot versions) were entered in the regression in step 1. Baseline motivation to quit, $\beta = 0.20$, $p = .007$, and the motivational interviewing chatbot (vs. confrontational counselling), $\beta = 0.14$, $p = .040$, significantly predicted UES score, together explaining 6.72% of the variance in UES score $F$ (8,191) = 1.72, $p = .096$, albeit not significant. In step 2, *acknowledgement* ($\beta$ *=0.005*, $p = .972$), *self-reflection* ($\beta = 0.29$, $p = .015$), and *motivational talk* ($\beta = 0.17$, $p = .041$) from the utterance features were entered, and the new model explained 12.9% of the variance in UES score, $F$ (11,188) = 2.53, $p = .005$. This step contributed to a small yet significant improvement of the model fit, $\Delta R^2 = 0.06$, $p = .003$. In step 3, *authentic*, *cognitive process*, *negative tone*, and *positive emotion* from the LIWC features were entered. Cognitive process ($\beta = 0.24$, $p = .002$) and negative tone ($\beta = 0.17$, $p = .020$) were significant predictors, and this step brought a bigger and significant increase ($\Delta R^2 = 0.08$, $p < .001$) to the model fit of the new model, $F$ (15,184) = 3.32, $p < .001$. In the last step, *number of utterances* and *average utterance length* were entered, and they did not appear to be significant predictors, nor did they contribute to a significant model fit increase. The results are presented in Table 5. In summary, words indicating active cognitive processes was found to be the strongest predictor of engagement, followed by motivational talk and overall negative tone.

## 5 DISCUSSION

This paper presented a study on textual cues and their relation to user engagement, both collected from a dataset of user-chatbot conversations. The primary objective is to explore whether and which of these cues can be used to infer and possibly predict user engagement. There is currently a lack of evaluation and validation of potential cues for engagement monitoring, Using a lens-based approach, we sought to address this gap by exploring the relationship between textual expressions and self-report user engagement. Ultimately, this exploration aims to contribute to a better understanding and measurement of user engagement, which can provide insights into the design and evaluation of human-chatbot interactions. In this section, we discuss implications of our findings.

### 5.1 Main findings and implications

We started with a rapid search of the literature, looking for various cues that have been proposed in the literature. Despite the limited body of literature directly pertaining to engagement cues, we compiled an inventory of cues utilized in analyzing human-chatbot interactions, The cues can be organized into three groups, ranging from thin-sliced utterance-level indicators (e.g., user complaints) to broader linguistic characteristics (e.g., the overall tone), and extending to more general interaction features (e.g., number of utterances). These cues serve as a lens through which user engagement is inferred. However, we found that the application of the lens perspective was primarily conducted with embodied agents, and the evaluation of textual cues was still in its early stage. A notable observation is that some studies operationalize engagement as dropout rates or the extent of usage [38, 57], while the experiential aspect of engagement received less consideration. We argue that dropout or continuous usage are, however, outcomes of engagement, rather than the process itself. The dynamic nature of engagement requires understanding changes in the experience over time, a challenge we contributed to by examining specifically on utterance-level cues. Our annotation process demonstrated initial feasibility of recognizing these cues from user utterances, as shown by a good interrater agreement. An important finding is that we had very few observations of several cues, such as complaints to the chatbot, expression of frustration or negative opinions about the topic. A potential explanation relates to the particular use case of the chatbot. The health counselling context in this study is non-task-oriented and led by the chatbot, wherein users did not have a specific request or a task to complete. In such scenarios, the chatbot may have been perceived more as an empathic helper with good intentions rather than a tool for completing a task [28]. This is in line with previous research suggesting that people in general tend to be more tolerant and exhibit positive responses (e.g., politeness) to empathic chatbots [65], particularly those characterized by a social or supportive intention [48].

Upon closer examination, regression analyses confirmed the usefulness of predicting engagement through multiple cues, offering practical implications for chatbot design and evaluation. Cognitive processes emerged to be the strongest predictor of engagement. This finding resonates with the theoretical conceptualization of engagement, where cognitive involvement is a key dimension of the experience [21]. Some studies have explored strategies to enhance cognitive involvement, such as employing quizzes [30], gamification [20], and asking reflective questions [42]. These strategies have proven effective in maintaining user engagement, providing

**Table 5: Results of hierarchical regression of cues predicting self-report engagement.**

| Predictor | $\beta$ | $R^2$ | $F$ | $\Delta R^2$ |
|---|---|---|---|---|
| **Step 1** | | 0.07 | 1.72 | |
| Chatbot version | 0.14* | | | |
| Gender | 0.05 | | | |
| Age | -0.04 | | | |
| Years of smoking | 0.04 | | | |
| Daily cigarette consumption | -0.09 | | | |
| Nicotine dependence level | 0.05 | | | |
| Baseline motivation to quit | 0.20** | | | |
| **Step 2** | | 0.13 | 2.53** | 0.06** |
| Chatbot version | -0.13 | | | |
| Gender | 0.08 | | | |
| Age | 0.01 | | | |
| Years of smoking | 0.02 | | | |
| Daily cigarette consumption | -0.10 | | | |
| Nicotine dependence level | 0.04 | | | |
| Baseline motivation to quit | 0.09 | | | |
| Acknowledgement | 0.00 | | | |
| Self-reflection | 0.29* | | | |
| Motivational talk | 0.17* | | | |
| **Step 3** | | 0.21 | 3.32*** | 0.08*** |
| Chatbot version | -0.05 | | | |
| Gender | 0.19 | | | |
| Age | -0.01 | | | |
| Years of smoking | -0.03 | | | |
| Daily cigarette consumption | -0.15 | | | |
| Nicotine dependence level | 0.10 | | | |
| Baseline motivation to quit | 0.08 | | | |
| Acknowledgement | 0.01 | | | |
| Self-reflection | 0.17 | | | |
| Motivational talk | 0.23** | | | |
| Authentic | 0.06 | | | |
| Cognitive process | 0.24** | | | |
| Negative tone | 0.16* | | | |
| Positive emotion | 0.09 | | | |
| **Step 4** | | 0.21 | 3.32*** | 0.00 |
| Chatbot version | -0.05 | | | |
| Gender | 0.21 | | | |
| age | -0.01 | | | |
| Years of smoking | -0.03 | | | |
| Daily cigarette consumption | -0.15 | | | |
| Nicotine dependence level | 0.10 | | | |
| Baseline motivation to quit | 0.08 | | | |
| Acknowledgement | 0.01 | | | |
| Self-reflection | 0.18 | | | |
| Motivational talk | 0.23* | | | |
| Authentic | 0.08 | | | |
| Cognitive process | 0.26** | | | |
| Negative tone | 0.17* | | | |
| Positive emotion | 0.08 | | | |
| Number of utterances | 0.02 | | | |
| Average utterance length | -0.05 | | | |

significance level: ***= $p < .001$, ** = $p < .01$, *= $p < .05$

further support for our findings. Furthermore, motivational talk (e.g., utterances in favor of quitting smoking) emerged as another predictor of engagement. When a user is motivated to discuss smoking cessation, they are naturally more likely to be engaged in the interaction. Conversely, active engagement can also enhance motivation for behavior change [47]. This reciprocal relationship between personal motivation and engagement highlights the need for more user-centered design. For instance, research in healthcare settings found that a chatbot using motivation-enhancing strategies was perceived as more engaging than a chatbot without those features [29]. Similarly, in the customer service domain, chatbots can enhance the overall customer experience by addressing users' goals and interests and providing personalized interactions [3]. These findings emphasize the importance of designing chatbot interactions adaptive to users' needs and motivations, ultimately leading to more effective engagement.

Especially noteworthy is the perhaps less intuitive finding that an overall negative tone turned out to be a positive predictor of engagement. This finding raises questions about the reliability of automated evaluation tools in capturing the nuanced meanings embedded within user utterances. While the overall tone was calculated automatically by the LIWC program, the semantic meaning of an utterance may differ from its mere linguistic expression. For instance, a user utterance "Smoking is bad for my health" has a negative tone but indicates motivations for behavior change. Such discrepancies between the lexical and the semantic dimensions of a user utterance highlights the importance of human judgement and contextual awareness in the annotation and analysis processes of user engagement. In text-based interactions, the absence of non-verbal cues poses additional challenges in interpretating users' intentions [62], calling for caution when relying on automated analysis tools. Human judgement plays a crucial role not only in user studies but also in chatbot system evaluation [40]. While automated approaches leverage labor and time efficiency, they have yielded mixed results, with some producing suboptimal predictions or results that deviate from ground truth [16, 57], emphasizing the need for nuanced interpretation of the contextual user input. Moreover, this phenomenon may be particularly prevalent in our specific dataset, where the negative consequences of smoking were frequently discussed. Therefore, we encourage further explorations using larger datasets from different domains to validate and expand upon our findings.

## 5.2 Future directions

A major bottleneck in evaluating thin-sliced engagement cues lies in the labor and time-intensive nature of manual annotation required for individual datasets. However, our demonstration of the feasibility of recognizing these cues from user utterances suggests promising areas for future research. Automating this process would be fruitful, since the ultimate goal of the recognition of engagement cues is to enable the chatbot system to monitor and maintain user engagement throughout interactions. Previous research in this area has shown that machine learning models were able to accurately predict dropout, but they struggle to predict more fine-grained levels of engagement [57], a limitation we aimed to address in this paper. As discussed above, human judgement and contextual understanding are essential in the annotation process. Future research is encouraged to integrate both human expertise and algorithmic approaches, ensuring validity and accuracy of predictions.

It should be noted that the final regression model showed that these cues only accounted for 21% of variance in engagement, suggesting that there remain unobserved factors that influence user engagement. One area worth of investigation relates to the interplay between chatbot performance and user reactions. In our study, we focused on user utterances only, without considering the content or quality of the chatbot's utterances. For instance, frequent error or inadequacies in the chatbot's responses could potentially lead to disengagement, even if users choose not to explicitly express their negative impression, as observed in our dataset. This was observed by some research [53], such that wrong intent recognition and off-topic responses can lead to poor user experience, and such non-lexical errors (unlike typos) can largely benefit from annotation approaches as outlined in this paper. Therefore, future research could consider including both users' and chatbots' utterances for a more comprehensive analysis of the interaction. This could involve error analyses to assess the frequency and nature of errors made by the chatbot and their potential relationship between user responses and engagement [4].

A main limitation of this work is that we used a single dataset from a specific domain. While we found several domain-agnostic cues in the literature (e.g., emoticon, user asking questions), they did not turn out to be predictive of user engagement. Our study presents an initial exploration into the relationship between textual cues and user engagement within the specific healthcare context, while its generalizability to other domains remains to be examined. For instance, cues expressing user request or negative responses might be more prevalent in customer service domain [15]. Therefore, future research is needed to validate these cues across different domains. Such cross-domain validation efforts can provide further insights into whether certain cues hold universal values or whether they are domain-specific, contributing to a more comprehensive understanding of user engagement in human-chatbot interactions. However, we suggest that our lens-based approach is generally applicable, and we hope to provide future research with a useful perspective.

# 6 CONCLUSION

To address the challenges in measuring and monitoring user engagement throughout interactions, researchers and practitioners often use behavior observations as cues to infer engagement. However, current evaluation and replication of these cues are lacking. This paper aims to contribute to a richer understanding of user engagement, textual cues, and how they relate. We collected a number of cues from the literature and demonstrated the feasibility of recognizing these cues in user utterances using a dataset of user-chatbot conversations. We found significant correlations between user engagement and several cues, including linguistic cues (e.g., words indicating cognitive processes), motivational cues (e.g., utterances indicating motivation to quit smoking), and interaction cues (e.g., the length of utterances). At a closer look, regression analyses showed that cognitive processes are the strongest predictor of

engagement, followed by motivational cues and negative tone. Our findings provide methodological and practical implications for the design and evaluation of chatbots. Future research is encouraged to validate and expand upon our findings in different domains.

## ACKNOWLEDGMENTS

This project is funded by the Dutch Research Council (NWO) with project number 406.DI.19.054.

## REFERENCES


[1] Alaa Ali Abd-Alrazaq, Asma Rababeh, Mohannad Alajlani, Bridgette M. Bewick, and Mowafa Househ. 2020. Effectiveness and safety of using chatbots to improve mental health: systematic review and meta-analysis. *Journal of medical Internet research* 22, 7 (2020), e16021.

[2] Abhishek Aggarwal, Cheuk Chi Tam, Dezhi Wu, Xiaoming Li, and Shan Qiao. 2023. Artificial Intelligence–Based Chatbots for Promoting Health Behavioral Changes: Systematic Review. *Journal of Medical Internet Research* 25, (2023), e40789.

[3] Tarek Ait Baha, Mohamed El Hajji, Youssef Es-Saady, and Hammou Fadili. 2023. The power of personalization: A systematic review of personality-adaptive chatbots. *SN Computer Science* 4, 5 (2023), 661.

[4] Deepali Aneja, Daniel McDuff, and Mary Czerwinski. 2020. Conversational Error Analysis in Human-Agent Interaction. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents* (*IVA '20*), October 19, 2020. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3383652.3423901

[5] Sandeep Avula, Gordon Chadwick, Jaime Arguello, and Robert Capra. 2018. SearchBots: User Engagement with ChatBots during Collaborative Search. In *Proceedings of the 2018 Conference on Human Information Interaction&Retrieval - CHIIR '18*, 2018. ACM Press, New Brunswick, NJ, USA, 52–61. https://doi.org/10.1145/3176349.3176380

[6] Erkan Basar, Divyaa Balaji, Linwei He, Iris Hendrickx, Emiel Krahmer, Gert-Jan de Bruijn, and Tibor Bosse. 2023. HyLECA: A Framework for Developing Hybrid Long-term Engaging Controlled Conversational Agents. In *Proceedings of the 5th International Conference on Conversational User Interfaces*, 2023. 1–5.

[7] Timothy Bickmore, Laura Pfeifer, and Daniel Schulman. 2011. Relational agents improve engagement and learning in science museum visitors. In *Intelligent Virtual Agents: 10th International Conference, IVA 2011, Reykjavik, Iceland, September 15-17, 2011. Proceedings 11*, 2011. Springer, 55–67.

[8] Timothy Bickmore, Daniel Schulman, and Langxuan Yin. 2010. Maintaining Engagement in Long-Term Interventions with Relational Agents. *Applied Artificial Intelligence* 24, 6 (July 2010), 648–666. https://doi.org/10.1080/08839514.2010.492259

[9] Stacey Birkett, Adam Galpin, Simon Cassidy, Lynne Marrow, and Sarah Norgate. 2011. How revealing are eye-movements for understanding web engagement in young children. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, May 07, 2011. ACM, Vancouver BC Canada, 2251–2256. https://doi.org/10.1145/1979742.1979900

[10] Ned Block. 1981. Psychologism and behaviorism. *The Philosophical Review* 90, 1 (1981), 5–43.

[11] Ryan L. Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W. Pennebaker. 2022. The development and psychometric properties of LIWC-22. *Austin, TX: University of Texas at Austin* (2022), 1–47.

[12] Egon Brunswik. 1956. *Perception and the representative design of psychological experiments*. University of California Press, Berkley.

[13] Jonathan Carlton, Andy Brown, Caroline Jay, and John Keane. 2021. Using Interaction Data to Predict Engagement with Interactive Media. In *Proceedings of the 29th ACM International Conference on Multimedia* (*MM '21*), October 17, 2021. Association for Computing Machinery, New York, NY, USA, 1258–1266. https://doi.org/10.1145/3474085.3475631

[14] Jacky Casas, Marc-Olivier Tricot, Omar Abou Khaled, Elena Mugellini, and Philippe Cudré-Mauroux. 2020. Trends & Methods in Chatbot Evaluation. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*, October 25, 2020. ACM, Virtual Event Netherlands, 280–286. https://doi.org/10.1145/3395035.3425319

[15] Ana Paula Chaves and Marco Aurelio Gerosa. 2021. How should my chatbot interact? A survey on social characteristics in human–chatbot interaction design. *International Journal of Human–Computer Interaction* 37, 8 (2021), 729–758.

[16] Jason Ingyu Choi, Ali Ahmadvand, and Eugene Agichtein. 2019. Offline and Online Satisfaction Prediction in Open-Domain Conversational Systems. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, November 03, 2019. ACM, Beijing China, 1281–1290. https://doi.org/10.1145/3357384.3358047

[17] Martin Colbert and Angela Boodoo. 2011. Does' Letting Go of the Words' increase engagement: A traffic study. In CHI'11 Extended Abstracts on Human Factors in Computing Systems. 655–667.

[18] M. Csikszentmihalyi. 1990. Flow. The Psychology of Optimal Experience. New York (HarperPerennial) 1990. (1990).

[19] Sara Dalzel-Job, Craig Nicol, and Jon Oberlander. 2008. Comparing behavioural and self-report measures of engagement with an embodied conversational agent: a first report on eye tracking in Second Life. In *Proceedings of the 2008 symposium on Eye tracking research & applications* (*ETRA '08*), March 26, 2008. Association for Computing Machinery, New York, NY, USA, 83–85. https://doi.org/10.1145/1344471.1344493

[20] Benner Dennis, Sofia Schöbel, Andreas Janson, and Jan Marco Leimeister. 2024. Engaging Minds–How Gamified Chatbots can Support and Motivate Learners in Digital Education. In *Hawaii International Conference on System Sciences (HICSS)*, 2024. .

[21] Kevin Doherty and Gavin Doherty. 2019. Engagement in HCI: Conception, Theory and Measurement. *ACM Comput. Surv.* 51, 5 (September 2019), 1–39. https://doi.org/10.1145/3234149

[22] Michelle Drouin, Susan Sprecher, Robert Nicola, and Taylor Perkins. 2022. Is chatting with a sophisticated chatbot as good as chatting online or FTF with a stranger? *Computers in Human Behavior* 128, (March 2022), 107100. https://doi.org/10.1016/j.chb.2021.107100

[23] Asbjørn Følstad, Marita Skjuve, and Petter Bae Brandtzaeg. 2019. Different chatbots for different purposes: towards a typology of chatbots to understand interaction design. In *Internet Science: INSCI 2018 International Workshops, St. Petersburg, Russia, October 24–26, 2018, Revised Selected Papers 5*, 2019. Springer, 145–156.

[24] Fotos Frangoudes, Marios Hadjiaros, Eirini C. Schiza, Maria Matsangidou, Olia Tsivitanidou, and Kleanthis Neokleous. 2021. An overview of the use of chatbots in medical and healthcare education. In *International Conference on Human-Computer Interaction*, 2021. Springer, 170–184.

[25] Hannah Gaffney, Warren Mansell, and Sara Tai. 2019. Conversational agents in the treatment of mental health problems: mixed-method systematic review. *JMIR mental health* 6, 10 (2019), e14166.

[26] Kai Guo, Yuchun Zhong, Danling Li, and Samuel Kai Wah Chu. 2023. Investigating students' engagement in chatbot-supported classroom debates. *Interactive Learning Environments* (2023), 1–17.

[27] Linwei He, Divyaa Balaji, Reinout W. Wiers, Marjolijn L. Antheunis, and Emiel Krahmer. 2023. Effectiveness and acceptability of conversational agents for smoking cessation: A systematic review and meta-analysis. *Nicotine and Tobacco Research* 25, 7 (2023), 1241–1250.

[28] He, Linwei, Erkan Basar, Emiel Krahmer, Reinout W. Wiers, and Marjolijn L. Antheunis. Effectiveness and user experience of a smoking cessation chatbot: A mixed-methods study comparing motivational interviewing and confrontational counselling.

[29] Linwei He, Erkan Basar, Reinout W. Wiers, Marjolijn L. Antheunis, and Emiel Krahmer. 2022. Can chatbots help to motivate smoking cessation? A study on the effectiveness of motivational interviewing on engagement and therapeutic alliance. *BMC Public Health* 22, 1 (2022), 726.

[30] Khe Foon Hew, Weijiao Huang, Jiahui Du, and Chengyuan Jia. 2021. Using chatbots in flipped learning online sessions: perceived usefulness and ease of use. In *Blended Learning: Re-thinking and Re-defining the Learning Process. 14th International Conference, ICBL 2021, Nagoya, Japan, August 10–13, 2021, Proceedings 14*, 2021. Springer, 164–175.

[31] Jennifer Hill, W. Randolph Ford, and Ingrid G. Farreras. 2015. Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. *Computers in Human Behavior* 49, (August 2015), 245–250. https://doi.org/10.1016/j.chb.2015.02.026

[32] Robert Ho. 2006. Handbook of univariate and multivariate data analysis and interpretation with SPSS. CRC press.

[33] Ryo Ishii and Yukiko I. Nakano. 2008. Estimating user's conversational engagement based on gaze behaviors. In *Intelligent Virtual Agents: 8th International Conference, IVA 2008, Tokyo, Japan, September 1-3, 2008. Proceedings 8*, 2008. Springer, 200–207.

[34] Carolina Islas Sedano, Verona Leendertz, Mikko Vinni, Erkki Sutinen, and Suria Ellis. 2013. Hypercontextualized learning games: Fantasy, motivation, and engagement in reality. *Simulation & Gaming* 44, 6 (2013), 821–845.

[35] Klaus Krippendorff. 2009. Testing the reliability of content analysis data. *The content analysis reader* (2009), 350–357.

[36] Alexander J. Kull, Marisabel Romero, and Lisa Monahan. 2021. How may I help you? Driving brand engagement through the warmth of an initial chatbot message. *Journal of Business Research* 135, (October 2021), 840–850. https://doi.org/10.1016/j.jbusres.2021.03.005

[37] Effie Lai-Chong Law, Asbjørn Følstad, and Nena Van As. 2022. Effects of Humanlikeness and Conversational Breakdown on Trust in Chatbots for Customer Service. In *Nordic Human-Computer Interaction Conference*, 2022. 1–13.

[38] Chi-Hsun Li, Ken Chen, and Yung-Ju Chang. 2019. When There is No Progress with a Task-Oriented Chatbot: A Conversation Analysis. In Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile

Devices and Services, October 2019. ACM, Taipei Taiwan, 1–6. https://doi.org/10.1145/3338286.3344407

[39] Weixin Liang, Kai-Hui Liang, and Zhou Yu. 2021. HERALD: An Annotation Efficient Method to Detect User Disengagement in Social Conversations. Retrieved September 21, 2023 from http://arxiv.org/abs/2106.00162

[40] Chia-Wei Liu, Ryan Lowe, Iulian V. Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. arXiv preprint arXiv:1603.08023 (2016).

[41] Yukiko I. Nakano and Ryo Ishii. 2010. Estimating user's engagement from eye-gaze behaviors in human-agent conversations. In Proceedings of the 15th international conference on Intelligent user interfaces, 2010. 139–148.

[42] Johanna Nurmi, Keegan Knittle, Todor Ginchev, Fida Khattak, Christopher Helf, Patrick Zwickl, Carmina Castellano-Tejedor, Pilar Lusilla-Palacios, Jose Costa-Requena, and Niklas Ravaja. 2020. Engaging users in the behavior change process with digitalized motivational interviewing and gamification: development and feasibility testing of the precious app. JMIR mHealth and uHealth 8, 1 (2020), e12884.

[43] Heather L. O'Brien, Paul Cairns, and Mark Hall. 2018. A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *International Journal of Human-Computer Studies* 112, (2018), 28–39.

[44] Heather L. O'Brien and Elaine G. Toms. 2008. What is user engagement? A conceptual framework for defining user engagement with technology. *Journal of the American society for Information Science and Technology* 59, 6 (2008), 938–955.

[45] Robert M. O'brien. 2007. A caution regarding rules of thumb for variance inflation factors. *Quality & quantity* 41, (2007), 673–690.

[46] Catharine Oertel, Ginevra Castellano, Mohamed Chetouani, Jauwairia Nasir, Mohammad Obaid, Catherine Pelachaud, and Christopher Peters. 2020. Engagement in human-agent interaction: An overview. *Frontiers in Robotics and AI* 7, (2020), 92.

[47] Olga Perski, Ann Blandford, Robert West, and Susan Michie. 2017. Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis. *Translational behavioral medicine* 7, 2 (2017), 254–267.

[48] Minjin Rheu, Yue Dai, Jingbo Meng, and Wei Peng. 2024. When a Chatbot Disappoints You: Expectancy Violation in Human-Chatbot Interaction in a Social Support Context. *Communication Research* (2024), 00936502231221669.

[49] Hanan Salam and Mohamed Chetouani. 2015. Engagement detection based on mutli-party cues for human robot interaction. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2015. IEEE, 341–347.

[50] Henrik Schoenau-Fog. 2011. Hooked!–evaluating engagement as continuation desire in interactive narratives. In *Interactive Storytelling: Fourth International Conference on Interactive Digital Storytelling, ICIDS 2011, Vancouver, Canada, November 28–1 December, 2011. Proceedings 4*, 2011. Springer, 219–230.

[51] Huma Shah, Kevin Warwick, Jordi Vallverdú, and Defeng Wu. 2016. Can machines talk? Comparison of Eliza with modern dialogue systems. *Computers in Human Behavior* 58, (May 2016), 278–295. https://doi.org/10.1016/j.chb.2016.01.004

[52] Annika Silvervarg and Arne Jonsson. Subjective and Objective Evaluation of Conversational Agents in Learning Environments for Young Teenagers.

[53] Marita Skjuve, Asbjørn Følstad, and Petter Bae Brandtzaeg. 2023. The User Experience of ChatGPT: Findings from a Questionnaire Study of Early Users. In *Proceedings of the 5th International Conference on Conversational User Interfaces*, July 19, 2023. ACM, Eindhoven Netherlands, 1–10. https://doi.org/10.1145/3571884.3597144

[54] Sinarwati Mohamad Suhaili, Naomie Salim, and Mohamad Nazim Jambli. 2021. Service chatbots: A systematic review. *Expert Systems with Applications* 184, (2021), 115461.

[55] S. Shyam Sundar, Saraswathi Bellur, Jeeyun Oh, Qian Xu, and Haiyan Jia. 2014. User experience of on-screen interaction techniques: An experimental investigation of clicking, sliding, zooming, hovering, dragging, and flipping. *Human–Computer Interaction* 29, 2 (2014), 109–152.

[56] Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology* 29, 1 (2010), 24–54.

[57] Ha Trinh, Ameneh Shamekhi, Everlyne Kimani, and Timothy W. Bickmore. 2018. Predicting User Engagement in Longitudinal Interventions with Virtual Agents. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, November 05, 2018. ACM, Sydney NSW Australia, 9–16. https://doi.org/10.1145/3267851.3267909

[58] B. Troy Frensley, Marc J. Stern, and Robert B. Powell. 2020. Does student enthusiasm equal learning? The mismatch between observed and self-reported student engagement and environmental literacy outcomes in a residential setting. *The Journal of Environmental Education* 51, 6 (November 2020), 449–461. https://doi.org/10.1080/00958964.2020.1727404

[59] Alexandria Katarina Vail, Joseph F. Grafsgaard, Joseph B. Wiggins, James C. Lester, and Kristy Elizabeth Boyer. 2014. Predicting Learning and Engagement in Tutorial Dialogue: A Personality-Based Model. In *Proceedings of the 16th International Conference on Multimodal Interaction*, November 12, 2014. ACM, Istanbul Turkey, 255–262. https://doi.org/10.1145/2663204.2663276

[60] Niels Van Berkel, Denzil Ferreira, and Vassilis Kostakos. 2017. The experience sampling method on mobile devices. *ACM Computing Surveys (CSUR)* 50, 6 (2017), 1–40.

[61] Anu Venkatesh, Chandra Khatri, Ashwin Ram, Fenfei Guo, Raefer Gabriel, Ashish Nagar, Rohit Prasad, Ming Cheng, Behnam Hedayatnia, Angeliki Metallinou, Rahul Goel, Shaohua Yang, and Anirudh Raju. 2018. On Evaluating and Comparing Open Domain Dialog Systems. Retrieved September 21, 2023 from http://arxiv.org/abs/1801.03625

[62] Marsha White and Steve M. Dorman. 2001. Receiving social support online: implications for health education. *Health Education Research* 16, 6 (December 2001), 693–707. https://doi.org/10.1093/her/16.6.693

[63] Zhou Yu, Vikram Ramanarayanan, Patrick Lange, and David Suendermann-Oeft. 2019. An Open-Source Dialog System with Real-Time Engagement Tracking for Job Interview Training Applications. In *Advanced Social Interaction with Agents: 8th International Workshop on Spoken Dialog Systems*, Maxine Eskenazi, Laurence Devillers and Joseph Mariani (eds.). Springer International Publishing, Cham, 199–207. https://doi.org/10.1007/978-3-319-92108-2_21

[64] Ran Zhao, Oscar J. Romero, and Alex Rudnicky. 2018. SOGO: A Social Intelligent Negotiation Dialogue System. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, November 05, 2018. ACM, Sydney NSW Australia, 239–246. https://doi.org/10.1145/3267851.3267880

[65] Qi Zhou, Bin Li, Lei Han, and Min Jou. 2023. Talking to a bot or a wall? How chatbots vs. human agents affect anticipated communication quality. *Computers in Human Behavior* 143, (June 2023), 107674. https://doi.org/10.1016/j.chb.2023.107674

[66] LWT Chatbots. Retrieved February 29, 2024 from https://lwt.cls.ru.nl/chat/home

# APPENDICESS

## A1 PAPER DETAILS FOR CUE EXTRACTION

Table A1 below provides information of papers from which we extracted the textual cues. This mainly include cues that were manually annotated; domain specific cues that are not necessarily within the chatbot context (e.g., change talk, sustain talk, and self-reflection) are not included in this table. Please note that this was not a systematic literature review; it rather served as a rapid search of relevant papers providing cues that we could test in our dataset.

**Table A1: Paper details for cue extraction.**

| Author years | Chatbot context | Textual cues | Cue level | Engagement measure |
|---|---|---|---|---|
| Liange et al., 2021 | Social chatbot | Complain bot repetition.<br>Complain bot ignoring user.<br>Complain bot misunderstanding.<br>Not understand bot.<br>Express frustration.<br>Show low interest.<br>Expression negative opinion about the interaction. | Utterance-level | No human involved; a machine learning model classifying engagement vs. disengagement |
| Trinh et al. 2018 | Health counselling | Interaction duration.<br>Number of utterances. | Interaction-level | Dropout rates |
| | | Average utterance length.<br>User asking questions.<br>User repeats | Utterance-level | Dropout rates |
| Vail et al., 2014 | Education | Acknowledgement | Utterance-level | UES score |
| Venkatesh et al., 2018 | Social chatbot | Number of dialogues turns.<br>Total interaction duration | Interaction-level | User rating (scale unknown) |
| Hill et al., 2015 | Social chatbot | Number of utterances | Interaction-level | Comparison with human-human interactions |
| | | Average utterance length.<br>Shorthand.<br>Emoticons. | Utterance-level | Comparison with human-human interactions |
| Li et al., 2019 | Customer service | Restatement. | Utterance-level | Drop out |
| Zhao et al., 2018 | Social debating chatbot | Total dialogue length.<br>Average utterance length | Interaction-level | User rating on attentiveness and perceived rapport |
| Drouin et al. 2022 | Social chatbot | LIWC variables.<br>Shorthand.<br>Emoticons. | Utterance-level | Comparison with human-human interactions |
| Kull et al. 2021 | Customer service | LIWC variables | Utterance-level | Not clear |

## A2 Correlation table between cues (that are not zero-inflated) and UES

**Table A2: Correlation matrix of textual cues and UES scores.**

| | L1 | L2 | L3 | L4 | L5 | L6 | L7 | L8 | L9 | U1 | U2 | U3 | U4 | U5 | U6 | I1 | I2 | I3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| L1 | | | | | | | | | | | | | | | | | | |
| L2 | 0.37 | | | | | | | | | | | | | | | | | |
| L3 | -0.37 | -0.34 | | | | | | | | | | | | | | | | |
| L4 | -0.48 | -0.34 | 0.54 | | | | | | | | | | | | | | | |
| L5 | 0.10 | 0.18 | -0.24 | -0.10 | | | | | | | | | | | | | | |
| L6 | -0.05 | -0.12 | 0.12 | 0.03 | -0.10 | | | | | | | | | | | | | |
| L7 | -0.03 | -0.09 | 0.05 | 0.08 | 0.36 | 0.13 | | | | | | | | | | | | |
| L8 | -0.06 | -0.04 | 0.13 | 0.02 | -0.04 | 0.78 | 0.08 | | | | | | | | | | | |
| L9 | 0.09 | 0.37 | -0.38 | -0.25 | 0.28 | -0.11 | -0.15 | -0.12 | | | | | | | | | | |
| U1 | -0.04 | 0.00 | 0.18 | 0.01 | -0.14 | 0.11 | 0.21 | 0.12 | -0.18 | | | | | | | | | |
| U2 | -0.22 | -0.15 | 0.20 | 0.12 | 0.01 | 0.22 | 0.32 | 0.22 | -0.14 | 0.35 | | | | | | | | |
| U3 | -0.24 | -0.07 | 0.25 | 0.14 | 0.07 | 0.19 | 0.30 | 0.19 | -0.06 | 0.31 | 0.80 | | | | | | | |
| U4 | -0.24 | -0.05 | 0.25 | 0.19 | 0.02 | 0.25 | 0.28 | 0.26 | -0.08 | 0.21 | 0.66 | 0.70 | | | | | | |
| U5 | -0.05 | 0.07 | 0.17 | 0.04 | 0.11 | 0.19 | 0.30 | 0.27 | -0.09 | 0.35 | 0.60 | 0.58 | 0.53 | | | | | |
| U6 | -0.30 | -0.34 | 0.22 | 0.26 | -0.26 | 0.16 | 0.03 | 0.07 | -0.16 | 0.06 | 0.20 | 0.17 | 0.13 | -0.28 | | | | |
| I1 | -0.14 | -0.09 | 0.36 | 0.12 | -0.04 | 0.15 | 0.29 | 0.22 | -0.17 | 0.54 | 0.66 | 0.61 | 0.50 | 0.56 | 0.20 | | | |
| I2 | -0.37 | -0.17 | 0.73 | 0.56 | -0.23 | 0.12 | 0.17 | 0.16 | -0.31 | 0.21 | 0.29 | 0.37 | 0.41 | 0.30 | 0.24 | 0.44 | | |
| I3 | -0.21 | -0.11 | 0.44 | 0.26 | -0.15 | 0.11 | 0.24 | 0.19 | -0.28 | 0.37 | 0.45 | 0.48 | 0.46 | 0.37 | 0.25 | 0.70 | 0.62 | |
| UES | 0.06 | 0.03 | 0.14 | 0.17 | 0.08 | 0.15 | 0.18 | 0.10 | -0.01 | -0.08 | 0.18 | 0.21 | 0.26 | 0.31 | -0.19 | 0.15 | 0.18 | 0.07 |

Color coding: light green: $p$<.05, medium-light green: $p$<.01; dark green: $p$<.001.

Code for cues: L1: Analytic; L2: Clout; L3: Authentic; L4: cognitive processes; L5: positive tone; L6: negative tone; L7: positive emotion; L8: negative emotion; L9: social processes; U1: restatement; U2: acknowledgement; U3: low-level self-reflection; U4: medium-level self-reflection; U5: change talk; U6: sustain talk; I1: number of utterances; I2: average utterance length; I3: total interaction duration.