

# Towards Mutual Theory of Mind in Human-AI Interaction: How Language Reflects What Students Perceive About a Virtual Teaching Assistant

Qiaosi Wang  
Georgia Institute of Technology  
Atlanta, GA, USA  
qswang@gatech.edu

Koustuv Saha  
Georgia Institute of Technology  
Atlanta, GA, USA  
koustuv.saha@gatech.edu

Eric Gregori  
Georgia Institute of Technology  
Atlanta, GA, USA  
egregori3@gatech.edu

David A. Joyner  
Georgia Institute of Technology  
Atlanta, GA, USA  
david.joyner@gatech.edu

Ashok K. Goel  
Georgia Institute of Technology  
Atlanta, GA, USA  
ashok.goel@cc.gatech.edu

## ABSTRACT

Building conversational agents that can conduct natural and prolonged conversations has been a major technical and design challenge, especially for community-facing conversational agents. We posit Mutual Theory of Mind as a theoretical framework to design for natural long-term human-AI interactions. From this perspective, we explore a community's perception of a question-answering conversational agent through self-reported surveys and computational linguistic approach in the context of online education. We first examine long-term temporal changes in students' perception of Jill Watson (JW), a virtual teaching assistant deployed in an online class discussion forum. We then explore the feasibility of inferring students' perceptions of JW through linguistic features extracted from student-JW dialogues. We find that students' perception of JW's anthropomorphism and intelligence changed significantly over time. Regression analyses reveal that linguistic verbosity, readability, sentiment, diversity, and adaptability reflect student perception of JW. We discuss implications for building adaptive community-facing conversational agents as long-term companions and designing towards Mutual Theory of Mind in human-AI interaction.

## CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; • **Human-centered computing** → **Natural language interfaces**; *Empirical studies in collaborative and social computing*; *Social media*; • **Applied computing** → *Psychology*.

## KEYWORDS

conversational agent, online community, human-AI interaction, theory of mind, language analysis, online education

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).  
CHI '21, May 8–13, 2021, Yokohama, Japan

© 2021 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 978-1-4503-8096-6/21/05...\$15.00  
<https://doi.org/10.1145/3411764.3445645>

## ACM Reference Format:

Qiaosi Wang, Koustuv Saha, Eric Gregori, David A. Joyner, and Ashok K. Goel. 2021. Towards Mutual Theory of Mind in Human-AI Interaction: How Language Reflects What Students Perceive About a Virtual Teaching Assistant. In *CHI Conference on Human Factors in Computing Systems (CHI '21)*, May 8–13, 2021, Yokohama, Japan. ACM, New York, NY, USA, 15 pages. <https://doi.org/10.1145/3411764.3445645>

## 1 INTRODUCTION

Conversational Agents (CAs)<sup>1</sup> are becoming increasingly integrated into various aspects of our lives, providing services across health-care, entertainment, retail, and education. While CAs are relatively successful in task-oriented interactions [82, 96], the initial promise of building CAs that can carry out natural and coherent conversations with users has largely remained unfulfilled due to both design and technical challenges [3, 18, 87]. This “gulf” between user expectation and experience with CAs [61] has led to constant user frustration, frequent conversation breakdowns, and eventual abandonment of CAs [3, 61, 98].

Conducting smooth conversations with users becomes even more crucial when CAs are deployed in online communities, especially those catering to vulnerable populations such as online health support groups [71] and student communities [95]. These community-facing CAs often serve as a critical part of the community to ensure smooth interactions among community members and provide long-term informational and emotional support. However, these community-facing CAs face two unique challenges: the need to carry out smooth dyadic interactions with individual community members, and the need to respond accordingly based on the community's shifting perceptions [53, 86]. In fact, the community-facing nature of the CA adds new complexity—each dyadic interaction with individual members is visible to other community members, which can not only change the community's perception of the CA, but can also impact other community members, i.e., unsatisfactory interaction with one individual might also frustrate others [42]. However, humans are able to gracefully conduct smooth interactions with each other and behave according to a community's

<sup>1</sup>Unless indicated otherwise, in this paper, we use CAs to refer specifically to disembodied, text-based conversational agents.

expectations and norms at the same time. This process is based on a uniquely humane characteristic called “*Theory of Mind*” [7, 12, 78].

Scholars posit that the **Theory of Mind (ToM) is a basic cognitive and social characteristic that enables us to make conjectures about each others’ minds through observable or latent behavioral and verbal cues** [6, 12, 37, 38, 94]. This characteristic spontaneously drives our understanding of how we perceive each other during social interactions. This enables us to employ social techniques such as adjusting our appearances and behaviors to align others’ perceptions about us based on our self-presentation [36]. In typical human-human interactions, having a **Mutual Theory of Mind (MToM), meaning all parties involved in the interactions possess the ToM**, builds a shared expectation of each other through behavioral feedback, helping us to maintain constructive and coherent conversations [36, 75]. MToM is increasingly used as a theoretical framework for the design of human-centered AI, such as robots, that can be perceived as more “natural” and intelligent during collaborations with human partners [26, 57, 59, 75].

While MToM is influencing the design of human-centered AI in task-oriented interactions, its role in informing the design of human-AI communicative interactions remains unexplored. Existing approaches to designing human-AI interactions also lack a theoretical framework and a unified design guideline to design human-centered CAs, especially in communicative interactions. Consequently, researchers and designers turn to traditional HCI design guidelines intended for Graphical User Interfaces, which is not always the optimal perspective to look at designing the interactions between humans and often anthropomorphized CAs [89]— researchers and designers face major obstacles in balancing unrealistically high user expectations [61] while providing an adequate amount of social cues to facilitate long-term natural interactions [56].

In analogy to human-human interactions, we propose *designing towards MToM as a theoretical framework to guide the design of adaptive community-facing CAs that can cater to users’ changing perceptions and needs*. **The first step towards building MToM in human-CA communications is thus equipping the CAs with an analog of ToM that can automatically identify user perceptions about the CAs**. With this ability, CAs would be able to monitor users’ changing perceptions and provide subtle behavioral cues accordingly to help users build a better mental model about CA’s capability. This would also help alleviate the current one-sided communication burden on users, who had to constantly adjust their mental model of the CA through an arbitrary trial-and-error process to elicit desired CA responses [4, 9].

Research has explored ways along the realm of identifying user perceptions of CAs to facilitate dyadic human-AI interactions, including examining an individual’s mental model of CAs in a variety of contexts [31, 54, 61]. These studies, most of which are qualitative in nature, are not only difficult to scale, but also lack directly feasible algorithmic outcomes that could be integrated into CA architecture to automatically recognize user perception about the CA. For community-facing CAs that are known to have fluid social roles in online communities [87], we presently lack a clear understanding of how community perception of CAs evolve over time, and whether the very dyadic interactions between humans and CAs in community settings reveal any signal related to user perceptions.

We thus note a gap in theory and practice in automatically and scalably understanding human perceptions of a community-facing CAs at both individual and collective level. Drawing on the dynamics of human-human interactions, this paper explores a first step towards designing for MToM in long-term human-CA interactions by examining the feasibility of building community-facing CAs’ ToM. Specifically, we target two research questions:

**RQ 1:** How does a community’s perception of a community-facing CA change over time?

**RQ 2:** How do linguistic markers of human-AI interaction reflect perception about the community-facing CA?

We examine these research questions within the context of online learning, where community-facing CAs are commonly seen to provide informational and social support to student communities [1, 92, 95]. We deployed a community-facing question-answering (QA)CA named **Jill Watson** [24, 34, 35] (JW for short) in an online class discussion forum to answer students’ questions for 10 weeks over the course of a semester. We collected students’ bi-weekly self-reported perceptions and conversations with JW for further analysis. We discuss changes in the student community’s long-term perception of JW and examine the relationship between self-reported student perceptions of JW and linguistic attributes of student-JW conversations such as verbosity, adaptability, diversity, and readability. Regression analyses between linguistic attributes and student perceptions of JW reveal insightful findings such as readability, sentiment, diversity and adaptability positively vary with desirable perceptions, whereas verbosity varies negatively.

Our contributions are three-fold: *First*, we propose MToM as the theoretical framework to design prolonged human-AI interaction within online communities. *Second*, our work provides a deeper understanding of how a community’s perception of a community-facing (QA)CA fluctuates longitudinally. *Third*, we provide empirical evidence on the potential of leveraging computational linguistic approach to infer community perception of a community-facing CA through accumulated public dyadic interactions within the community context. We discuss the implications of our work in designing adaptive community-facing (QA)CAs through theory-driven computational linguistic approaches, where our ultimate goal concerns building natural, long-term human-AI interactions.

**Privacy, Ethics, and Disclosure.** We are committed to ensuring the privacy of students’ data used in this study. This study was approved by the Institutional Review Board (IRB) at Georgia Tech. We collected the survey and discussion forum data (limited to only student-JW interactions) by seeking student consent and the data was anonymized. We offered extra credits to students for filling out each survey, and bonus extra credits if they completed at least five out of the six surveys. This work was in collaboration with the class instructor and we took measures to avoid coercion. The maximum number of extra credits students could earn by participation was less than 1% of the total grade, and these extra credits could also be earned in other ways as part of the standard class structure. We clarified to the students that survey responses would not be shared with the instructor, and would not have any impact on grades.

## 2 BACKGROUND

In this section, we provide an overview of ToM and its application in existing human-AI interaction research. We then discuss related work that explores user perception of the CA to facilitate human-AI interactions and highlight the potential of leveraging language analysis to improve human-AI interaction.

### 2.1 Theory of Mind in Human-AI Interaction

ToM, our ability to make suppositions of other’s minds through behavioral cues, is fundamental to many human social and cognitive behaviors, especially our ability to collaboratively accomplish goal-oriented tasks and our ability to perform smooth, natural communications with others [6, 7]. For example, building shared plans and goals is fundamental to collaborative task-completion—ToM enables us to recognize and mitigate each other’s plans and goals to work together [5, 6]; intentional communication is the basis for smooth communication—ToM enables us to understand the interlocutor has a belief or knowledge that can potentially be altered and thus allows us to compose our messages accordingly [5, 6]. Without ToM, our ability to naturally interact with others can be severely impaired and render our ability to perceive and produce language less meaningful [5–7, 12]. ToM has been extensively studied and continues to be a leading influence in many fields such as cognitive science, developmental psychology, and autism research [5, 7, 47, 78].

Over the years, researchers have recognized the importance of designing human-centered robots with ToM to facilitate collaborations within human-robot teams for goal-oriented tasks. Specifically, ToM has been intentionally built in as an individual module of the system architecture to help robots monitor world state as well as the human state [25], construct simulation of hypothetical cognitive models of the human partner to account for human behaviors that deviate from original plans [44, 79], and help robots to build mental models about user beliefs, plans and goals [43, 52]. Robots built with ToM have demonstrated positive outcomes in team operations [25, 44], collaborative decision-making [39], and are perceived to be more natural and intelligent [59].

However, ToM has not been explored as a built-in characteristic that can potentially enable CAs to communicate naturally with humans. Given its fundamental role in human interactions and success in human-robot interactions so far, we posit that the role of ToM in designing human-centered CAs during communicative interactions with humans should be further examined. Building CAs with ToM is the first step towards designing for *Mutual* ToM in human-CA interactions. By *Mutual* ToM we meant to not only explore how to help users build a better understanding and mental model of CAs (e.g., explainable AI), but also to consider how to help CAs build and iterate on a comprehensive mental model of the user. In this paper, we present our initial exploration towards building MToM in human-CA communications by examining the feasibility of building CA’s ToM using automatic language analysis of human-CA conversations to understand user perception of CAs.

### 2.2 User Perception of Conversational Agent

Our perception of CAs determines how we interact with them and thus plays a crucial role in guiding the design of human-centered CAs. People’s perception of CAs is a multifaceted concept. Prior

research has explored people’s mental model of CAs in various settings—in a cooperative game setting, people’s mental model of the CA could include global behavior, knowledge distribution, and local behavior [31]; people’s perception of a recommendation agent consists of trust, credibility, and satisfaction [10]. People’s perception of CAs is instrumental in guiding how they interact with CAs [31] and thus serves as a precursor to their expectation of CA behavior. Prior research has suggested that users tend to hold high expectations of CAs [61] and thus prone to encounter frequent conversation breakdowns, which can ultimately lead users to abandon the CA [58, 98]. Recognizing user perception of CAs and provide appropriate feedback to help users revise their perceptions is thus critical in building smooth human-CA interactions [9, 31, 41].

Understanding user perceptions of CAs becomes even more important in the context of online learning where CAs are being increasingly used to play a critical part in students’ learning experience. A variety of CAs has been designed to offer learning and social support to students. These include Intelligent Tutoring Systems that provide individualized learning support for students [1, 46], community-facing CAs that provide synchronous online lectures [95] and those help to build social connections among online learners [92]. However, while these CAs have been found to be efficient in facilitating students’ learning outcome, how students perceive these CAs in the online classroom is barely explored—students’ perceptions of the CA could influence students’ interaction experience with the CA and thus potentially impacting their online learning experience [46].

For community-facing CAs specifically, understanding the community’s perception of the CA is not only important to ensure smooth dyadic interactions within the community, but also crucial to design community-facing CAs as long-term companions. Many research suggested that community-facing CAs have shifting social roles and thus perceived by the community differently on a frequent basis. For example, Seering et al. [88] found that CA’s social role within an online learning community shifted from being the “dependent” to the “peer” over time within the community; Kim et al. [53] also highlighted that CAs can potentially shift their role from encouraging participation to “social organizer” as the community dynamic evolves over time. Yet more nuanced assessment of community perception of CAs over time is needed for CAs to behave appropriately based on their changing social roles.

However, the majority of the research on human perception of CAs uses qualitative methods to either identify the various dimensions of user perception of the agent [31, 67] or provides one-time assessments of user perception after interacting with the agent [30, 41, 51, 55, 88, 98]. These work, while offered valuable insights and guidelines for the design of human-centered CAs, are difficult to operationalize and integrate into the CA’s system architecture. Post-hoc analysis of people’s perception of CA is also less effective in capturing the nuance and fluidity of people’s perception of CA overtime during interactions.

Current work thus seeks to examine the long-term variations of community perception of a community-facing CA. In order to operationalize community perception of the CA, we also explore the feasibility of capturing community perceptions of CA automatically by leveraging computational linguistics.



### 2.3 Language Analysis to Improve Human-AI Interaction

Language plays an instrumental role in all kinds of interactions, yet it is the most paramount component and often the only component of human-CA interactions. Designing MToM in human-CA interactions thus heavily depends on the information both humans and CAs convey through their textual responses.

To help CAs understand the user's mental model of CAs in a timely yet non-intrusive manner, it is important to explore the feasibility of inferring user's perceptions of CAs from language. Prior research suggested the potential of leveraging linguistic cues to indicate people's perception of CAs during human-CA interactions. Researchers have inferred users' emotions towards the agent [90], personality traits [62], signs of conversation breakdowns [58, 99], politeness [20] from conversational cues. Yet whether a user's holistic perception of the CA could be constructed through linguistic characteristics extracted from conversations remains unexplored.

On the other hand, more research is needed to explore how can CA's language convey its capability to the user and help guide users to revise their mental model about the CA. Whereas humans tend to leverage various social techniques such as altering appearances and manners to make certain impressions [36], embodied virtual agents and robots can change facial expressions or physical behaviors [73], voice-based conversation assistants can change the tone of their voice [11], text-based language is the only way for disembodied, text-based CAs to convey their capabilities to the user. CA responses have been shown to influence user's repair strategies [9] and foster engaging conversations [14, 33]. CA's language in responses thus should be leveraged to help users construct a better mental model of the CA [9].

With the ultimate goal of building MToM in human-CA interactions by leveraging language analysis, in this work, we first begin by examining the feasibility of inferring user perception of the CA through linguistic features extracted from human-CA conversations. Automatically construct user's mental models of CAs during conversations can enable CAs to accurately predict user behavior and provide appropriate responses to guide users to tune their mental models in order to facilitate smoother human-CA conversations.

## 3 STUDY DESIGN

Current study seeks to understand longitudinal changes of community perception of a community-facing CA and the feasibility of leveraging linguistic markers to infer user perceptions of a community-facing CA. To explore these questions, we deployed a community-facing (QA)CA named "Jill Watson (JW)" in an online class discussion forum to answer students' class logistic questions throughout the semester. We then collected students' bi-weekly perceptions of JW and extracted linguistic characteristics from student-JW conversations over the course of the semester (see Figure 1 for detailed study design). We selected our survey measures and linguistic features with the goal of ultimately building CA's ToM—survey measures were designed to gauge students' perception of JW from three dimensions: anthropomorphism, intelligence, and likeability; linguistic characteristics were suggested by prior literature to have the potential of reflecting people's perception of the CA. We discuss them in detail in the following sections.

Current study took place in an online human-computer interaction class offered through the Online Master of Science in Computer Science (OMSCS) program at Georgia Tech. This class had 376 students enrolled at the end of the semester. Based on a separate standard class survey that asked for students' demographic information at the beginning of the semester ( $n=389$ , note that some students dropped the class before the end of the semester), 299 students self-identified as male (76.86%), 87 students self-identified as female (22.37%), three students did not report their gender (0.77%). Students were spread out across various age groups: 61 students were between 18 to 24 (15.68%), 236 students were between 25 to 34 (60.67%), 65 students were between 35 to 44 (16.71%), 22 students were between 45 to 54 (5.66%), two students were between 55 to 64 (0.51%), and two students were above 65 years old (0.51%), one student did not report their age. In terms of highest education degree obtained, 311 students reported Bachelor's Degree (79.95%), 56 students reported Master's Degree (14.40%), 14 students reported Doctoral Degree (e.g. PhD., Ed.D.) (3.60%), and seven students reported Professional Degree (e.g. M.D., J.D.) (1.80%), one student did not report their highest degree obtained.

### 3.1 Design and Implementation of JW

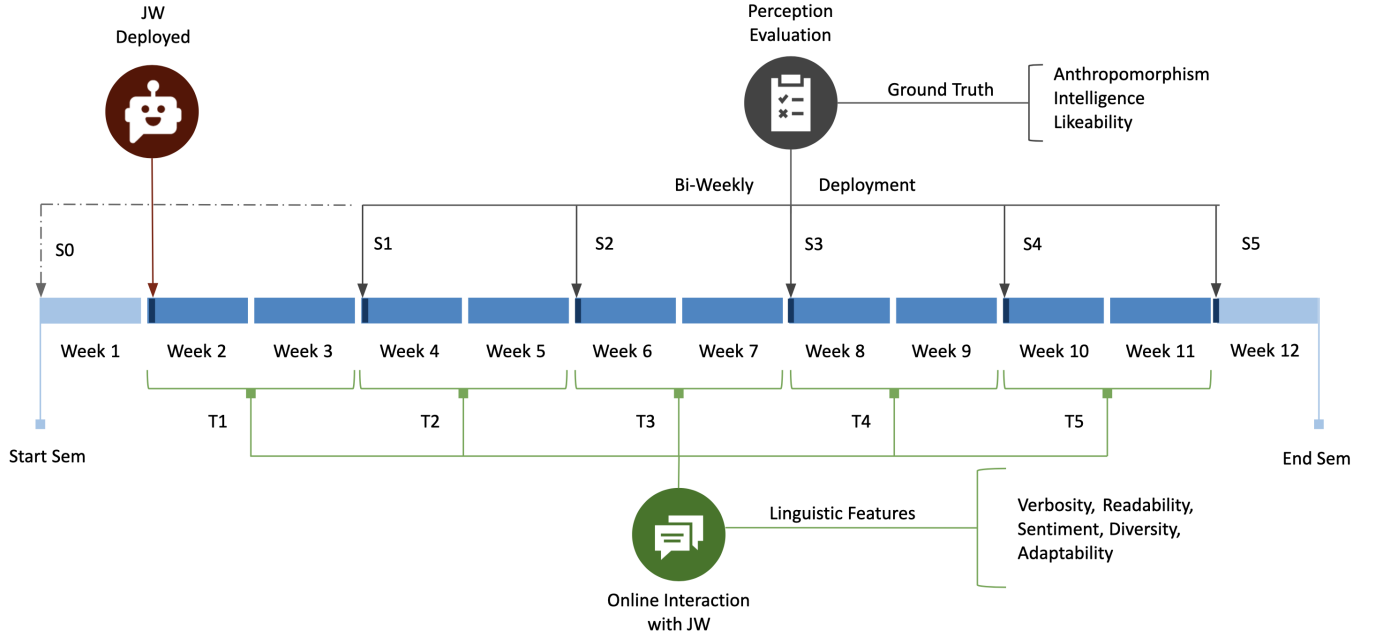
JW is an ML-based question-answering CA designed to answer students' questions about class logistics. It uses three machine learning models with each model being trained with the same data. When a user asks a question, the question is passed to all three models. The final output of the models is used to select a pre-programmed response (greetings + relevant information in the syllabus). The models were trained using training questions generated from a knowledge base. The knowledge base was created using a syllabus ontology and the course syllabus. JW thus *cannot* learn from outside information (student responses or feedback) over time. Implementation details of similar previous versions of JW can be found in Goel and Polepeddi [35].

We deployed JW on the class discussion forum at the beginning of the second week (Figure 1). JW was only active on dedicated "JW threads" where JW read and provided responses to each post to only questions posted in those threads. Students were encouraged to post their class-related questions on this thread if they wanted an answer from JW. To keep students engaged throughout the semester, we posted a new "JW thread" every week on the discussion forum and encouraged students to keep asking questions to JW. Table 1 shows a list of example question-answer pairs between the students and JW on the class discussion forum.

Throughout our study, we *intentionally did not specify JW's working mechanism or capabilities to the students* so that the information would not bias the students' perception of JW. The students were only told that JW was a virtual agent who could answer their questions about the class. JW's working mechanism and implementation were only revealed after all the survey data was collected.

## 4 RQ1: EXAMINING CHANGES IN STUDENT PERCEPTIONS ABOUT JW

To explore changes in students' perceptions of JW throughout the semester, we deployed six bi-weekly surveys (See Appendix Figure 3 for the adapted survey instrument) for students to self-report their

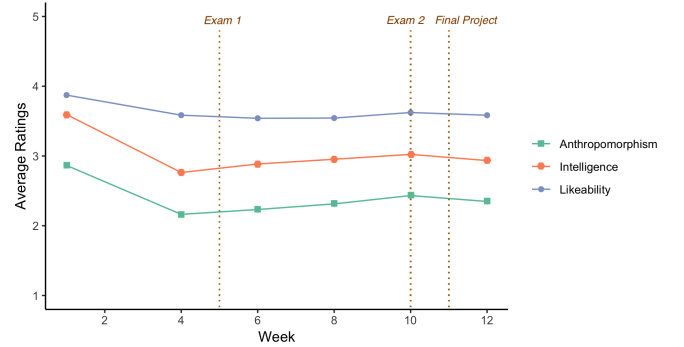


**Figure 1: Study design and timeline.** S0-S5 represents the survey data. T1-T5 represents our division of class discussion forum data based on the survey distribution timeline. In the regression analysis, we used survey data as ground truth to tag student interaction with JW in each time frame. For instance, we used S1 to tag forum data from T1, S2 to tag T2, and so on.

perceptions of JW. Inspired by the MToM theoretical framework, we intentionally selected perception metrics that could capture students’ holistic social perceptions of JW and potentially reflect long-term changes in perceptions of JW, instead of the commonly measured post-hoc perceptions of CA functionalities (e.g., accuracy or correctness of response). In particular, we adapted a validated survey instrument measuring user perception of robots in human-robot interactions [8], also previously applied in human-CA interaction settings [50]. In our specific setting of student-JW interactions, our surveys inquired students to self-report their perceptions about JW along three dimensions: 1) anthropomorphism, 2) intelligence, and 3) likeability. In addition, we also asked the students to report how/if they interacted with JW in the past two weeks (e.g., read other students’ interactions with JW, posted questions to JW).

**Data.** We started with an initial total dataset of 1513 responses from S<sub>0</sub> to S<sub>5</sub>. We consolidated all our responses to build our final dataset that included all valid, complete responses from students who indicated that they interacted with JW by either reading through other students’ interactions or posting questions to JW. We ended up with a total of 1132 responses from S<sub>0</sub> to S<sub>5</sub> ( $N_{S_0} = 260$ ,  $N_{S_1} = 201$ ,  $N_{S_2} = 171$ ,  $N_{S_3} = 171$ ,  $N_{S_4} = 164$ ,  $N_{S_5} = 165$ ).

Our analyses did not include S<sub>0</sub> survey results that indicate students’ expectation of JW prior to actual interactions as we are more interested in examining long-term changes in student perception *after* at least some initial interaction. It is also well-established in the literature that people often have unrealistically high expectations for CAs [61, 98]. Our findings replicate this similar pattern from prior literature, that students’ perception *decreased* compared



**Figure 2: Student perceptions of JW over time.** To provide more context, the plot marks the due dates of Exam 1, Exam 2, and Final Project. Note that students in this class also have weekly written assignments.

to their initial perception (or expectation) as per S<sub>0</sub>, as is revealed in Figure 2, which plots the aggregated community perception about JW over the course of the semester.

#### 4.1 Changes in Perceived Anthropomorphism, Intelligence, and Likeability of JW

Next, to understand if student perception of JW changed significantly after initial interactions, we performed Kruskal-Wallis test [64]

**Table 1: Examples of question-answer pairs during students' interactions with JW throughout the semester on the class discussion forum thread.**

Week	Student Question	JW Answer
Week 4	What is the next assignment?	Hi, happy to answer your question! There are ten written assignments in the course: five P assignments and five M assignments. The P assignments are Principles assignments and cover the material covered in Unit 2. Each Principles assignment asks you to [...]. The M assignments are Methods assignments and cover the material in Unit 3. Each Methods assignment asks for [...].
Week 4	How were you programmed?	Hi, happy to answer your question! Sorry I don't understand your question. Could you rephrase your question by composing a new followup discussion?
Week 6	Hi Jill Watson, is the exam going to cover the concepts we applied in the assignments?	Hi, happy to answer your question! Here are some questions that I think you might be asking. Please compose a new followup discussion with one of the questions here: "Can I work ahead?", "What is the JDF specification?" "Can we look at online resources?"
Week 8	What's the most efficient way to earn participation points?	Hi, happy to answer your question! HCI is a deeply collaborative field, and there is no better way to learn than to interact with your peers, to participate in one another's usability studies, and to see the variety of approaches taken by your classmates to the class's assignments. Thus, participation credit may be earned in one of three ways: by completing peer reviews, by participating in one another's surveys and interviews, and by participating in discussions on Piazza. The following are the ways you may earn participation points [...]
Week 10	When is the mid-course survey due?	Hi, happy to answer your question! Please complete this survey by the end of week [link to mid-course survey]. You receive 1 participation points for completing one of the four course surveys. For more details, see the participation policy [...]
Week 12	Hi Jill Watson, could you predict the percentage of students who would get an "A" at the end of class?	Hi, happy to answer your question! Please complete this survey by the end of week [link to mid-course survey]. You receive 1 participation points for completing one of the four course surveys. For more details, see the participation policy [...]

on students' self-reported perception of JW from  $S_1$  to  $S_5$ . Kruskal-Wallis test is a non-parametric, omnibus test, which we used because our data is not normally distributed based on the results of a Shapiro-Wilk normality test ( $p < 0.001$ ). We then conducted further post-hoc pairwise comparison to examine differences between each bi-weekly perception report. Dunn Test result shows significant differences in perceived anthropomorphism between  $S_1$  and  $S_4$ :  $z = -2.82, p = 0.02$ , and significant differences in perceived intelligence between  $S_1$  and  $S_4$ :  $z = -3.26, p = 0.01$ . We reported the detail test results and effect sizes in Table 2.

**4.1.1 Anthropomorphism.** Anthropomorphism is the attribution of human characteristics to non-human objects such as computers and CAs. Anthropomorphism is a widely studied yet highly debatable design characteristic of CA—on one hand, intentionally building CAs with more humanlike attributes can improve user trust [13, 32], make the CA more approachable and ease user interactions [54, 97]; on the other hand, the famous "Uncanny Valley" effect [66] indicates that highly anthropomorphized CA could evoke people's negative feelings towards the CA [17] as well as setting unrealistic user expectations on CA's capabilities [61]. Changes in perceived anthropomorphism over time is thus an important quality to investigate as it could significantly affect people's expectation of the CA and thus influence trust-building and long-term human-agent relationship [23]. Kruskal-Wallis test found students' self-reported perceived anthropomorphism after initial interaction with JW *changed significantly* over time from  $S_1$  through  $S_5$ :  $\chi^2(4) = 9.55, p < 0.05$ .

**Table 2: Summary of comparison in students' bi-weekly perceptions of JW. We report Kruskal-Wallis test results for each perception metrics from  $S_1$  to  $S_5$ , the posthoc pairwise comparison  $z$  statistic (Dunn Test), and effect size (Cohen's  $d$ ).  $p$ -values are reported after Bonferroni correction (\*  $p < 0.05$ , \*\*  $p < 0.01$ ).**

Measure	Anthropomorphism		Intelligence		Likeability	
	$z$	$d$	$z$	$d$	$z$	$d$
$S_1$ and $S_2$	-0.60	0.08	-1.63	0.16	0.67	0.06
$S_1$ and $S_3$	-1.47	0.17	-2.32	0.25	0.69	0.06
$S_1$ and $S_4$	-2.82*	0.31	-3.26**	0.33	-0.59	0.05
$S_1$ and $S_5$	-1.88	0.21	-2.13	0.22	0.04	0.00
$S_2$ and $S_3$	-0.83	0.10	-0.66	0.09	0.02	0.00
$S_2$ and $S_4$	-2.14	0.22	-1.59	0.18	-1.20	0.11
$S_2$ and $S_5$	-1.23	0.13	-0.49	0.06	-0.60	0.06
$S_3$ and $S_4$	-1.32	0.13	-0.93	0.09	-1.22	0.11
$S_3$ and $S_5$	-0.41	0.04	0.16	0.02	-0.62	0.06
$S_4$ and $S_5$	0.90	0.10	1.09	0.11	0.60	0.05
Kruskal-Wallis	$\chi^2(4) = 9.55 **$		$\chi^2(4) = 11.81*$		$\chi^2(4) = 2.09$	

Post-hoc pairwise comparison found  $S_2$  and  $S_5$  differ significantly:  $z = -2.82, p < 0.05$ . This indicates that CAs' perceived humanlikeness by the community can vary over time, even when the agent has zero learning ability and adaptability.

**4.1.2 Intelligence.** Intelligence refers to the perceived level of intelligence of the CA by the community, in other words, how much users perceive the CA as an intelligent being. Even though building artificially "intelligent" machines has been a unfulfilled promise

due to various technical and feasibility challenges [8, 87], users tend to expect their CAs to be “smart” [98], thus creating a gap between user expectation and CA’s true intelligence. CA’s knowledge is also one of the key components identified in people’s mental model of CA [31]. Therefore, perceived intelligence plays an important role in how people perceive, evaluate, and interact with the agent. However, it is unclear whether people’s perception about the CA’s intelligence change over time. In this study, Kruskal-Wallis test found that the perceived intelligence of JW changed significantly from  $S_1$  to  $S_5$ :  $\chi^2(4) = 11.811, p < 0.05$ , specifically, post-hoc pairwise comparison shows that perceived intelligence reported in  $S_2$  and  $S_5$  differ significantly:  $z = -3.26, p < 0.01$ . This highlights a CA’s perceived intelligence is an important attribute to consider when building long-term human-AI relationships.

**4.1.3 Likeability.** Likeability refers to how likeable the interlocutor is perceived by others. In human interactions, likeability has been suggested to induce positive affect, increase persuasiveness, and foster favorable perceptions [76, 80]. Since people often treat computers as social actors [8, 70], perceived likeability is a potential factor that could influence long-term relationship-building. Kruskal-Wallis test could not find statistically significant changes in students’ self-reported likeability of JW over time:  $\chi^2(4) = 2.0947, p = 0.72$ . This result could be attributed to the fact that positive first impression in human interactions typically plays a crucial role in long-term likeability [81]. Another reason could be that students’ initial perception of JW remains the same over time since JW was intentionally designed to be a basic CA without learning ability.

**4.1.4 Correlation Between Perception Measures.** We also conducted Spearman correlation test, a non-parametric correlation test, to examine the relationship between these three perception measures. Spearman correlation results show that perceived anthropomorphism and intelligence have a strong positive relationship ( $r_s = (0.74), p < 0.001$ ), intelligence and likeability have a moderately strong positive relationship ( $r_s = (0.62), p < 0.001$ ), and anthropomorphism and likeability have a low positive correlation ( $r_s = (0.51), p < 0.001$ ). This result suggests that even though the three measures of perception are considered to be somewhat independent [8], that may not be the case in our data. That is, according to our data, students’ perception of desirability along the three measures are in similar direction, a general increasing trend in one would likely convey in a general increasing trend in the other two.

**4.1.5 Summary and Interpretation.** Through analyzing students’ bi-weekly self-report of their perception of JW, we conclude that JW’s perceived anthropomorphism and intelligence significantly changed over time, but perceived likeability did not significantly vary in the long run. Our findings help us understand how community perceptions of a community-facing CAs change. This bears implications on designing community-facing CAs to be able to adapt to community’s changing perceptions of the CA in the long run. We also found the three measures of self-reported perceptions to be inter-correlated, shedding light that these measures may not be very disentangled (or independent) in users’ mental models.

## 5 RQ2: EXAMINING THE LANGUAGE OF STUDENT-JW INTERACTIONS

In this section, we examine the relationship between how the students perceived and linguistically interacted with JW. To do this, we collected the conversation logs between students and JW from all the weekly question and answering threads on the public discussion forum and then extracted linguistic features for further data analysis. With the goal of exploring the feasibility of building a ToM for CAs, the linguistic measures were chosen due to their known potential in reflecting users’ holistic perceptions of CAs, which we refer to relevant research and describe in more details in the following sections. We also discuss the findings and implications for designing human-CA interactions.

### 5.1 Inferring Student Perception from Linguistic Features

First, we link students’ linguistic interactions with JW in a block of time with their immediate next self-reported perception about JW as ground-truth. For example, if a student made multiple posts to JW from week 4 to week 6 ( $T_2$ ) and reported their perception of JW in week 6 ( $S_2$ ), then for this student, we derive language features of  $T_2$  to understand their self-reported perception in  $S_2$ . Such an approach enables us to examine if the linguistic interaction between a student and JW in a block of time can predict how they would perceive the agent immediately at the end of that time block. This leads us to a total of 551 pairs of linguistic interactions and self-reported perceptions with  $N(T_1) = 157, N(T_2) = 86, N(T_3) = 126, N(T_4) = 96, N(T_5) = 86$ .

Next, we build linear regression models. Linear regression is known to help interpret conditionally monotone relationships with the dependent variable [22]. In particular, we build three linear regression models where each model uses one of the three perception measures as the dependent variable. We draw on prior research to derive a variety of linguistic attributes (features) from the language interactions which include verbosity, readability, sentiment, diversity, and adaptability [27, 84]. We use these linguistic features as independent variables in the models. As both perception and linguistic interactions could be a function of time, we include an ordinal variable of the week of the datapoint as a covariate in the models. Further, we control our models with an individual’s baseline language use, particularly the baseline average number of words computed over all the posts made by the same individual. Equation 1 describes our linear regression models, where  $\mathcal{P}$  refers to the measures of anthropomorphism, intelligence, and likeability.

$$\mathcal{P} \sim \text{Baseline} + \text{Week} + \text{Verbosity} + \text{Readability} + \text{Sentiment} + \text{Diversity} + \text{Adaptability} \quad (1)$$

**Summary of Models.** Our linear regression models reveal significance with  $R^2(\text{Anth.}) = 0.85, R^2(\text{Intel.}) = 0.93, R^2(\text{Like.}) = 0.95$ ; all with  $p < 0.001$ . Table 3 summarizes the coefficients of each dependent variable. First, we note the statistical significance of the control variables, *week* and *baseline word use*. We find that people who are more expressive are more likely to have a positive perception of the agent on all three perception measures. We find verbosity to be negatively associate with each measure of perception, while adaptability, diversity, and readability positively associate



with student perception of JW. Next, we describe our motivation, hypothesis, operationalization, and observation for each of our linguistic features below.

## 5.2 Linguistic Features: Motivation, Operationalization, and Observations

**5.2.1 Verbosity.** In human-human conversations, we tend to use shorter and less complex sentences when talking to a kid from sixth-grade versus when talking to an adult co-worker [68]. The verbosity of conversational language we produce thus depends on our mental model of how intelligent we perceive our interlocutor to be, which will drive the way we communicate our cognitive planning and execution of thoughts to others [27]. Translating from human-human to human-CA conversational settings, verbosity may vary on the basis of how intelligent and human-like we perceive the CA to be [45]. Hill et al. found that humans use less verbose and less complicated vocabulary when communicating with CAs, as compared to human-human conversations [45]. Further, the human-likeness of a CA could be judged based on the length of words used [60]. Someone who perceives a CA to be more human-like or more intelligent would likely use more verbose language. Accordingly in our setting, *we hypothesize that greater verbosity is associated with a more positive perception of JW.*

Drawing on prior work [45, 84], we use two measures to describe the verbosity of students' posts: 1) *length* and 2) *linguistic complexity*. We operationalize *length* as the number of unique words per post, and *complexity* as the average length of words per sentence [27, 84].

Our regression model (Table 3) suggests that both verbosity attributes show negative coefficients with all the perception measures along with statistical significance. This *rejects our hypothesis*. Contrary to prior research and popular belief [28, 45, 60], our findings suggest that students who used more number of unique words per post or more complex language tended to perceive JW as less human-like, less intelligent, and less likeable. We construe that more verbose and complex language could plausibly cause the CA to fail in providing supportive or efficacious responses, leading to undesirable CA perception.

**5.2.2 Readability.** Readability refers to the level of ease readers can comprehend a given text [63]. Psycholinguistic literature values readability to be a key indicator of people's cognitive behavior, and prior work has adapted this measure to understand conversational patterns in online communities [27, 84, 85]. While this measure has not been studied in the context of human-AI interactions, from the perspective of MToM, the readability of students' questions posted to JW can convey their perception of JW's text-comprehension ability. Therefore, we examine readability to understand students' interaction with JW. However, considering an analogy from human-human to human-AI conversations, we hypothesize that, *higher readability is an indicator of a more positive perception about the CA.*

To capture the readability of students' posts to JW, we calculate the Coleman-Liau Index (CLI). CLI is a readability assessment that approximates a minimum U.S. grade level required to understand a block of text, and is calculated using the formula:  $CLI = 0.0588L - 0.296S - 15.8$ , in which  $L$  is the average number of letters per 100 words and  $S$  is the average number of sentences per 100 words [77].

Our regression model shows that readability is positively associated with all three dimensions of student's perception of JW with statistical significance: anthropomorphism (2.33), perceived intelligence (2.41), and likeability (3.00). This result *supports our hypothesis*, suggesting that readability is a strong predictor of students' perception and positively varies with perception. This could be associated with an underlying intricacy that the more readable the question is, the more successful the CA response is, and the more satisfied (or positively perceiving) the users are.

**5.2.3 Sentiment.** During human-CA conversations, the emotion we convey through our language is often a manifestation of whether CA's perceived performance matches our expectations of the CA [61]. In fact, sentiment analysis has been used to detect customer satisfaction with customer service chatbots and yielded positive results [29]. Besides the perceived likeability of the CA, sentiment in the language is also positively associated with the perceived naturalness of the human-CA interactions [45, 72]. While there is a lack of evidence on how sentiment in wording can be associated with perceived intelligence, intelligence is one of the key desired characteristics people expect from a CA [54]. Therefore, *we hypothesize that sentiment in students' questions posted is positively associated with a positive perception of JW.*

To measure the sentiment of each post to JW, we used the VADER sentiment analysis model [48], which is a rule-based sentiment analysis model that provides numerical scores ranging from -1 (extreme negative) to +1 (extreme positive).

Our regression model (Table 3) shows a lack of evidence to support our hypothesis in the case of anthropomorphism, but a statistically significant support for hypothesis related to perceived intelligence (0.69) and likeability (0.64) with positive coefficients. The current study setting that JW was deployed in is considered a formal academic environment and thus themed discussion related to coursework is more common. We believe in settings where the affective language is much more prevalent (e.g., on online Reddit communities), sentiment might play a strong role in reflecting people's perception of a community-facing CA.

**5.2.4 Linguistic Diversity.** Depending on our perception of the interlocutor, the linguistic (and topical) diversity of our language could vary, i.e., the diversity of the conversation topics or the richness of language used. Linguistic diversity has been suggested to correlate with perceived intelligence during human-human interactions [68]. In human-CA interactions, when the CA behaves in a more natural and authentic way, users also tend to employ a richer set of language, conveying positive attitudes towards the CA [72]. Therefore, *we hypothesize that the greater the linguistic diversity is, the more positive students perceive JW.*

We draw on prior work [2, 84] to obtain linguistic diversity, and use word embeddings for this purpose. Word embeddings represent words as vectors in a higher dimensional latent space, where lexico-semantically similar words tend to have vectors that are closer [21, 65, 74]. In our case, for each post to JW, we first obtain its word embedding representation in 300-dimensional latent lexico-semantic vector space using pre-trained word embeddings [65]. We then compute the average cosine distance from the centroid of all the posts by the same user in each two-week period before



**Table 3: Coefficients of linear regression between students’ perception (as dependent variable) and language based measures of interaction with JW (as independent variables). Purple bars represent the magnitude of positive coefficients, and Golden bars represent the magnitude of negative coefficients. .  $p < 0.1$ , \*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$ .**

Measure	Anthropomorphism		Intelligence		Likeability	
	Coeff.	p	Coeff.	p	Coeff.	p
Baseline Avg. Num. Words	0.15	***	0.16	***	0.13	***
Week	0.06	***	0.06	***	0.03	**
Verbosity						
Num. Unique Words	-3.34	**	-3.37	**	-3.91	*
Complexity	-1.33	***	-1.82	***	-2.00	***
Readability	2.33	***	2.41	***	3.00	***
Sentiment	0.10		0.69	**	0.64	***
Linguistic Diversity	0.17	***	0.09		0.20	
Linguistic Adaptability	1.02	**	1.53	***	2.55	***
Adjusted R <sup>2</sup>	0.85	***	0.93	***	0.95	***

corresponding surveys. This operationalizes our measure of lexico-semantic diversity of each student’s post to JW.

According to our regression model, we find a lack of support for our hypothesis for perceived intelligence and likeability, whereas, a statistically significant support for our hypothesis on anthropomorphism which shows a positive coefficient (0.17). This finding adds some support to previous work on human-CA interaction that suggested positive association between high lexico-semantic diversity and perceived human-likeness of the CA [72]. Contradictory to observations related to human-human interactions [68], our observations suggest that people’s linguistic diversity does not necessarily indicate how intelligent one perceives an agent to be.

**5.2.5 Adaptability.** As humans, we tend to adapt to each other’s language use during conversations due to our inherent desire to avoid awkwardness in social situations [36]. Prior research suggested that people often mindlessly apply social rules and etiquette to computers [69], it is thus possible that we also adapt our language when conversing with a CA. In fact, prior work suggests that we are able to adapt our speech pattern accordingly based on whether the interlocutor is a human or a CA [45], suggesting that adaptability of our speech pattern could be an indicator of our perception of interlocutor’s intelligence, human-likeness, as well as likeability. Human users are more likely to build desirable perceptions about a CA if CA response is adapted and customized to human questions, as opposed to templated responses (e.g., “Thank You”, “Sorry”) [84]. Therefore, *We hypothesize that adaptability is positively associated to perceived anthropomorphism, likeability, and intelligence.*

Motivated by Saha and Sharma’s approach [84], we measure adaptability as the lexico-semantic similarity between each question-response pairs of student-JW interactions, operationalized as the cosine similarity of word embedding representations of the questions and responses. As in the case of diversity, we use 300-dimensional word embedding space [65].

Our regression model indicates that adaptability positively associates with anthropomorphism (1.02), intelligence (1.53), and likeability (2.55), all with statistical significance. This supports our hypothesis, and aligns with prior research on how people employ different speech patterns depending on if the interlocutor is a CA or a human [45]. Our observations suggest that adaptability is a valid

predictor of the perceptions of JW. We construe that if students receive adaptable responses, they are more likely to perceive JW as more human-like, likeable, and intelligent.

**5.2.6 Summary and Interpretations.** We examine the relationship between linguistic features of student-JW conversations and student perception of JW through regression analysis. We find that verbosity negatively associates with student perception of JW, whereas readability, sentiment, diversity, and adaptability positively associate with anthropomorphism, intelligence, and likeability. Our findings suggest the potential to extract linguistic features to measure community perceptions of CA during conversation, and thus enable the CA to constantly understand and provide desirable responses that match with user perception. It is important to note that the relationship between linguistic measures and three measures of student perception of JW is of the same degree and direction.

## 6 DISCUSSION

Our findings provide empirical evidence on the long-term variations in a community’s perception of a community-facing CA as well as the feasibility of inferring user perceptions of the CA through linguistic features extracted from the human-CA dialogue. Specifically, we found the student community’s perception of JW’s anthropomorphism and intelligence changed significantly over time, yet perceived likeability did not change significantly. Our regression analyses reveal that linguistic features such as verbosity, readability, sentiment, diversity, and adaptability are valid indicators of the community’s perceptions of JW. Based on these findings, we first discuss the implications of leveraging language analysis to facilitate human-AI interactions. Then, we present the challenges and opportunities for designing adaptive community-facing CAs. We also discuss the technical and design implications for human-CA communications and how future work can extend our findings towards building MToM in human-AI interactions.

### 6.1 Language Analysis to Design Human-AI Interactions

Our work demonstrates that leveraging linguistic features extracted from human-CA conversations has the potential to improve human-CA interactions. This technique, if properly integrated into the CA

design, would fulfill the promise of building truly “conversational” agents. Our findings indicate that language analysis can be used to automatically infer a community’s perception of a community-facing CA. This opens up the potential of using language analysis to design CAs that can automatically identify the user’s mental model of the CA, which allows the CAs to provide subtle hints in responses to guide the user in adjusting their mental model of the CA for a continuous and efficacious conversation.

In our study, even though JW is a question-answering(QA) CA designed to only fulfill students’ basic informational needs, we could infer student perceptions through language features extracted from these simple QA dialogues. Our findings resonate with prior work that also revealed the potential of using language analysis on question-answering conversational data between users and QA agents to infer conversation breakdowns [58]. We believe that in more sophisticated conversational settings where the human-CA interactions go beyond basic informational needs, and interactions that involve multimodal data (e.g., voice and visual communications), one can extract more nuanced descriptions of user perceptions about CAs. This would lead us to draw insights that can facilitate constructive and consistent human-CA dialogue.

We also note that student-JW interactions were situated in a much more controlled environment compared to many possible settings for human-CA interactions. For instance, the discussions in the online course forum are *supposed* to be thematically coherent about course work. Additionally, students are expected to self-present in a desirable and civil fashion—there are various online and offline norms and conventions that people tend to follow in academic settings [40]. On the other hand, discussions on a general-purpose online community (e.g., Reddit), including those which are moderated, can not only have diverse and deviant discussions but can also include informal languages [15]. These kinds of data can add noise to automated language models, and it opens up more research opportunities to examine how language in general-purpose online communities reflect the individual and collective perception about a community-facing CA.

Besides helping CAs understand how they are perceived by the users during interactions, language can also potentially indicate user preferences about the CA in a particular context and thus inform future design of CAs. For example, in our regression analyses, linguistic measures such as sentiment and diversity reflect similar directionality (see Table 3) among the correlation between the three perception measures (section 4.3)—we find a positive association between JW’s perceived intelligence and likeability, but weak correlation between anthropomorphism and likeability. In particular, sentiment extracted from the student-JW conversation is significantly associated with both intelligence and likeability, yet not significantly associated with perceived anthropomorphism. It is thus worth considering whether human-likeness is a more important factor to consider comparing to an agent’s intelligence demonstrated through providing informational support when designing virtual teaching assistants like JW. This finding also provides more evidence to the long-standing debate of whether CAs should be designed as humanlike as possible [16, 17], suggesting that user’s preference of whether CAs should be humanlike is highly dependent on CA’s role and use contexts.

## 6.2 Designing for Adaptive Community-Facing Conversational Agents

Prior work proposed seven social roles that community-facing CAs could serve within online human communities [87] yet how to quickly detect and measure people’s perceptions and expectations of how the CA should behave when serving different social roles remained unexplored. Our work opens up the opportunity to operationalize the desired social roles of community-facing CAs in terms of specific dimensions of CA perceptions. For example, when CA serves as a social organizer to help community members build social connections, the community could expect the CA to behave more humanlike and more likeable instead of more intelligent. These expectations could potentially be identified and monitored through linguistic cues, as demonstrated by our work. This operationalization can help community-facing CAs quickly identify the community’s expectations and produce behaviors that are better aligned with their perceived social roles within the community.

While prior research suggested community-facing CAs’ shifting social roles over time within online communities [51, 87, 88], our examination of long-term changes in the student community’s perception about JW provides empirical evidence on the specific variations in the community’s perception of the agent. Our findings indicate that community-facing CA’s perceived anthropomorphism and intelligence are more nuanced and fluid characteristics and thus require more frequent assessment for the CAs to adjust their behaviors within the community accordingly. JW’s perceived likeability did not change significantly in our study, suggesting that designers could have more leeway in monitoring CA’s perceived likeability. However, the reasoning behind JW’s stable perceived likeability within the community requires further examination—it could be because long-term likeability perception is highly dependent on first-impression [8], or it could be a result of JW’s stable performance over the semester due to its lack of learning ability.

One foreseeable challenge when designing adaptive community-facing CAs using linguistic cues to construct user perception of the CA is to distinguish the intention of each message—whether the user asked a genuine question or just trying to game the system; or whether the user’s reply was intended for the CA or other community members. While people employ strategies such as changing appearances to manage their self-presentation in daily lives [36], people also manage their self-presentation through linguistic cues on public online platforms, depending on the perceived audience [19, 49, 83]. For community-facing CAs, every dyadic human-CA interaction is visible to other community members as well. People thus might take advantage of this opportunity to not only gain support from the CA but also to modulate their responses to help manage their self-presentation within the community. For example, people might intentionally limit their emotional expression through language so that they don’t appear “stupid” for thinking a CA could interpret the emotional elements in the language [40]; or people might purposefully reply with questions that can help them appear more humorous than to receive a correct answer from the CA. There are several occurrences of this in our study when students ask JW questions that are clearly out of scope for JW, such as “*What is the meaning of life?*” or “*What is your favorite character in Game of Thrones?*”.

### 6.3 Towards Mutual Theory of Mind in Human-AI Interaction

With the ultimate goal of building MToM in Human-AI interactions, our study explored the feasibility of building a CA's ToM by operationalizing and identifying user perceptions of the CA through linguistic cues. From the lens of MToM, understanding the perception of each other during interactions, similar to human-human interaction, acts as the cognitive foundation of human-AI interactions [6]. Drawing upon human-human interactions that rely on all kinds of social signals conveyed through languages, to improve the accuracy of CA's ToM, future research should explore how other conversational cues could potentially be combined with the linguistic features we investigated to provide more context and accuracy in understanding people's perceptions of the CA. For example, identifying conversation breakdowns through conversational cues [58] or monitoring user emotions and satisfactions during interactions [29, 93] can be combined with the identified user perception of the CA. A potential implication is to keep the user perception identified through language analysis as a constant state, while other conversational cues can be used to trigger state change for the user's perception of the CA, and thus adapt the CA's behavior constantly.

While our work focuses on helping CAs to understand user's perceptions of the CAs through language analysis, we want to emphasize that *any type of communication is a two-way interaction*. To achieve *Mutual* ToM in human-CA interactions, it is also crucial to explore techniques that can help users understand CA's perceptions of their goals, intentions, skills, etc. for users to correct CA's perception in a timely fashion to achieve desirable human-CA interaction outcome. By proposing MToM as the theoretical framework to guide the design of human-CA interactions, our work motivates future research to further explore how to help users understand CA's perceptions of them. This direction could include examination of how and when the CA can offer their perception of the users in a way that is intuitive and easy-to-understand for the users, but also natural enough to maintain the authenticity of the conversation. This could include techniques that are currently being explored by researchers in explainable AI to help users have a sufficient understanding of how their textual responses would be parsed by the CA to extract perceptions about their goals and intentions [91].

Echoing with prior work's suggestion of understanding human-human social interactions as a way to improve human-AI interactions [33], we highlight the importance of leveraging interdisciplinary theoretical frameworks to offer new design perspectives on human-CA interactions. This work combines theories drawn from cognitive science and social science on human-human interactions—human social interactions are largely about impression management [36], which is dependent on the uniquely human cognitive ability of ToM [6, 12]. This enables us to rethink the design of human-AI interactions. We thus highlight the importance of borrowing theoretical frameworks from fields such as anthropology, cognitive science, social psychology, etc. to offer new design perspectives on human-AI interactions.

### 6.4 Limitations and Future Work

Our work has some limitations. Our results might not be transferable when human-CA interaction takes place in private dyadic interaction contexts. This work investigates the feasibility of inferring student perception of a community-facing CA through linguistic features extracted from dyadic human-agent interaction on a public discussion forum. Student perception and interaction with the agent thus might be biased by other students' interactions with the agent on the public forum, which we point out as a unique challenge to design for community-facing CAs that carry out dyadic interactions within human communities. Future research aimed at designing adaptive CAs in dyadic interactions could replicate the current study in one-to-one human-CA interactions.

Our work took a formative step towards understanding people's perception of a CA through linguistic features. Our findings are correlational and we cannot make causal claims. Future work that accounts for unobserved confounds can lead to better insights into human-AI perceptions and interactions. We also recognize more qualitative or mixed-methods approaches are needed to gain deeper insights into people's reasoning and intention behind their linguistic behaviors when conversing with a CA. For example, in our study, students could be intentionally testing if JW learned anything from their previous questions by posting the exact same questions from previous JW threads; or students might be frustrated by JW's learning ability and thus intentionally post difficult questions on the public thread — there is no way to evaluate this quantitatively, and future qualitative research could shed light on this issue.

To quantify student's perception of JW, we used a standardized measure taken from human-robot interaction that includes anthropomorphism, intelligence, and likeability [8]. However, the measurement we adopted does not suggest that these are, or should be, the standard dimensions of user perceptions of CA— in fact, prior research already suggested that there are different interpretations of how users build their mental models of CAs [10, 31]. We are, however, hopeful that language analysis can reveal the different dimensions of people's perceptions about CAs during interactions. Future research should replicate the current study using different measurements of the user's mental model about CA to provide more evidence on the potential of language analysis.

## 7 CONCLUSION

This paper posited Mutual Theory of Mind as the theoretical framework for designing adaptive community-facing conversational agents (CAs) as long-term companions. Guided by this framework, we examined the long-term changes of community perception of CA, and measured the feasibility of inferring perceptions through linguistic cues. We deployed a community-facing CA, JW, a virtual teaching assistant designed to answer students' logistical questions about the class in an online class public discussion forum. Driven by our understanding of Theory of Mind, we measured students' perception of JW in terms of perceived anthropomorphism, intelligence, and likeability. We found statistically significant long-term changes in student community's perception of JW in terms of anthropomorphism and intelligence. Then, we extracted theory-driven language features from student-JW interactions over the course of the semester. Regression analyses revealed that linguistic features such



as verbosity, diversity, adaptability, and readability explain students' perception of JW. We discussed the potential of leveraging language analysis to fulfill the promise of designing truly "conversational" agents, including the design implications of building adaptive community-facing CAs that can cater to community's shifting perceptions of the CA, and the theoretical implications of applying Mutual Theory of Mind as a design framework in facilitating human-AI interactions. We believe this research can inspire future work to leverage interdisciplinary theories to rethink human-CA interactions.

## ACKNOWLEDGMENTS

We thank Vedant Das Swain, Dong Whi Yoo, Adriana Alvarado Garcia, Scott Appling and the anonymous reviewers for their help and feedback. This work was funded through internal grants from Georgia Tech and Georgia Tech's College of Computing.

## REFERENCES

- [1] Vincent AWM Alven and Kenneth R Koedinger. 2002. An effective metacognitive strategy: Learning by doing and explaining with a computer-based cognitive tutor. *Cognitive science* 26, 2 (2002), 147–179.
- [2] Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *TACL* (2016).
- [3] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.
- [4] Zahra Ashktorab, Mohit Jain, Q. Vera Liao, and Justin D. Weisz. 2019. Resilient Chatbots: Repair Strategy Preferences for Conversational Breakdowns. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*. ACM Press, New York, New York, USA, 1–12. <https://doi.org/10.1145/3290605.3300484>
- [5] Simon Baron-Cohen. 1997. *Mindblindness: An essay on autism and theory of mind*. MIT press.
- [6] Simon Baron-cohen. 1999. Evolution of a Theory of Mind? In *The Descent of Mind: Psychological Perspectives on Hominid Evolution*. Oxford University Press, 1–31.
- [7] Simon Baron-Cohen, Alan M Leslie, Uta Frith, et al. 1985. Does the autistic child have a "theory of mind". *Cognition* 21, 1 (1985), 37–46.
- [8] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics* 1, 1 (1 2009), 71–81. <https://doi.org/10.1007/s12369-008-0001-3>
- [9] Erin Benetateu, Olivia K Richards, Mingrui Zhang, Julie A Kientz, Jason Yip, and Alexis Hiniker. 2019. Communication breakdowns between families and Alexa. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.
- [10] Emilie Bigras, Marc Antoine Jutras, Sylvain Sénécal, Pierre Majorique Léger, Marc Fredette, Chrystel Black, Nicolas Robitaille, Karine Grande, and Christian Hudon. 2018. Working with a recommendation agent: How recommendation presentation influences users' perceptions and behaviors. *Conference on Human Factors in Computing Systems - Proceedings* 2018-April (2018), 1–6. <https://doi.org/10.1145/3170427.3188639>
- [11] Michael Braun, Anja Mainz, Ronée Chadowitz, Bastian Pfleging, and Florian Alt. 2019. At your service: Designing voice assistant personalities to improve automotive user interfaces. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–11.
- [12] Peter Carruthers and Peter K Smith. 1996. *Theories of theories of mind*. Cambridge University Press.
- [13] Justine Cassell and Timothy Bickmore. 2000. External manifestations of trustworthiness in the interface. *Commun. ACM* 43, 12 (2000), 50–56.
- [14] Justine Cassell and Timothy Bickmore. 2003. Negotiated collusion: Modeling social language and its relationship effects in intelligent agents. *User Modelling and User-Adapted Interaction* 13, 1-2 (2003), 89–132. <https://doi.org/10.1023/A:1024026532471>
- [15] Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. 2017. The Bag of Communities: Identifying Abusive Behavior Online with Preexisting Internet Data. In *Proc. CHI*.
- [16] Dasom Choi, Daehyun Kwak, Minji Cho, and Sangsu Lee. 2020. "Nobody Speaks that Fast!" An Empirical Study of Speech Rate in Conversational Agents for People with Vision Impairments. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–13. <https://doi.org/10.1145/3313831.3376569>
- [17] Leon Ciechanowski, Aleksandra Przeglaska, Mikolaj Magnuski, and Peter Gloor. 2019. In the shades of the uncanny valley: An experimental study of human-chatbot interaction. *Future Generation Computer Systems* 92 (2019), 539–548.
- [18] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade, and Benjamin R. Cowan. 2019. What makes a good conversation? Challenges in designing truly conversational agents. *Conference on Human Factors in Computing Systems - Proceedings* (2019), 1–12. <https://doi.org/10.1145/3290605.3300705>
- [19] Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan Dumais. 2011. Mark my words! Linguistic style accommodation in social media. In *Proceedings of the 20th international conference on World wide web*. 745–754.
- [20] Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. *arXiv preprint arXiv:1306.6078* (2013).
- [21] Vedant Das Swain, Koustuv Saha, Manikanta D Reddy, Hemang Rajvanshy, Gregory D Abowd, and Munmun De Choudhury. 2020. Modeling Organizational Culture with Workplace Experiences Shared on Glassdoor. In *CHI*.
- [22] Robyn M Dawes and Bernard Corrigan. 1974. Linear models in decision making. *Psychological bulletin* 81, 2 (1974), 95.
- [23] Ewart J De Visser, Samuel S Monfort, Ryan McKendrick, Melissa AB Smith, Patrick E McKnight, Frank Krueger, and Raja Parasuraman. 2016. Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied* 22, 3 (2016), 331.
- [24] Chris Dede, John Richards, and Bror Saxberg. 2018. *Learning Engineering for Online Education: Theoretical Contexts and Design-based Examples*. Routledge.
- [25] Sandra Devin and Rachid Alami. 2016. An implemented theory of mind to improve human-robot shared plans execution. *ACM/IEEE International Conference on Human-Robot Interaction* 2016-April (2016), 319–326. <https://doi.org/10.1109/HRI.2016.7451768>
- [26] Bobbie Eicher, Kathryn Cunningham, Sydni Peterson Marissa Gonzales, and Ashok Goel. 2017. Toward mutual theory of mind as a foundation for co-creation. In *International Conference on Computational Creativity, Co-Creation Workshop*.
- [27] Sindhu Kiranmai Ernala, Asra F Rizvi, Michael L Birnbaum, John M Kane, and Munmun De Choudhury. 2017. Linguistic markers indicating therapeutic outcomes of social media disclosures of schizophrenia. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–27.
- [28] Hao Fang, Hao Cheng, Maarten Sap, Elizabeth Clark, Ari Holtzman, Yejin Choi, Noah A Smith, and Mari Ostendorf. 2018. Sounding board: A user-centric and content-driven social chatbot. *arXiv preprint arXiv:1804.10202* (2018).
- [29] Jasper Feine, Stefan Morana, and Ulrich Gnewuch. 2019. Measuring Service Encounter Satisfaction with Customer Service Chatbots using Sentiment Analysis. *Proceedings of the 14th International Conference on Wirtschaftsinformatik* December (2019), 0–11.
- [30] Radhika Garg and Subhasree Sengupta. 2020. Conversational Technologies for In-home Learning: Using Co-Design to Understand Children's and Parents' Perspectives. (2020), 1–13. <https://doi.org/10.1145/3313831.3376631>
- [31] Katy Ilonka Gero, Zahra Ashktorab, Casey Dugan, Qian Pan, James Johnson, Werner Geyer, Maria Ruiz, Sarah Miller, David R. Millen, Murray Campbell, Sadhana Kumaravel, and Wei Zhang. 2020. Mental Models of AI Agents in a Cooperative Game Setting. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–12. <https://doi.org/10.1145/3313831.3376316>
- [32] Eun Go and S Shyam Sundar. 2019. Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Computers in Human Behavior* 97 (2019), 304–316.
- [33] Rachel Gockley, Allison Bruce, Jodi Forlizzi, Marek Michalowski, Anne Mundell, Stephanie Rosenthal, Brennan Sellner, Reid Simmons, Kevin Snipes, Alan C Schultz, et al. 2005. Designing robots for long-term social interaction. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 1338–1343.
- [34] Ashok Goel. 2020. AI-Powered Learning: Making Education Accessible, Affordable, and Achievable. *arXiv preprint arXiv:2006.01908* (2020).
- [35] Ashok K Goel and Lalith Polepeddi. 2016. *Jill Watson: A virtual teaching assistant for online education*. Technical Report. Georgia Institute of Technology.
- [36] Erving Goffman. 1978. *The Presentation of Self in Everyday Life*. London: Harmondsworth.
- [37] Alvin I Goldman et al. 2012. Theory of mind. *The Oxford handbook of philosophy of cognitive science* 1 (2012).
- [38] Alison Gopnik and Henry M Wellman. 1992. Why the child's theory of mind really is a theory. (1992).
- [39] O Can Görür, Benjamin Rosman, and Guy Hoffman. 2017. Toward Integrating Theory of Mind into Adaptive Decision-Making of Social Robots to Understand Human Intention. In *Workshop on the Role of Intentions in Human-Robot Interaction at the International Conference on Human-Robot Interactions*. Vienna, Austria.



- [40] Pamela Grimm. 2010. Social desirability bias. *Wiley international encyclopedia of marketing* (2010).
- [41] Shivashankar Halan, Brent Rossen, Michael Crary, and Benjamin Lok. 2012. Constructionism of virtual humans to improve perceptions of conversational partners. (2012), 2387. <https://doi.org/10.1145/2212776.2223807>
- [42] Jeffrey T Hancock, Kailyn Gee, Kevin Ciaccio, and Jennifer Mae-Hwah Lin. 2008. I'm sad you're sad: emotional contagion in CMC. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*. 295–298.
- [43] Maaïke Harbers, Karel Van Den Bosch, and John Jules Meyer. 2009. Modeling agents with a theory of mind. *Proceedings - 2009 IEEE/WIC/ACM International Conference on Intelligent Agent Technology, IAT 2009 2* (2009), 217–224. <https://doi.org/10.1109/WI-IAT.2009.153>
- [44] Laura M. Hiatt, Anthony M. Harrison, and J. Gregory Trafton. 2011. Accommodating human variability in human-robot teams through theory of mind. *IJCAI International Joint Conference on Artificial Intelligence* (2011), 2066–2071. <https://doi.org/10.5591/978-1-57735-516-8/IJCAI11-345>
- [45] 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 (2015), 245–250. <https://doi.org/10.1016/j.chb.2015.02.026>
- [46] Kenneth Holstein, Bruce M McLaren, and Vincent Alevan. 2019. Designing for complementarity: Teacher and student needs for orchestration support in ai-enhanced classrooms. In *International Conference on Artificial Intelligence in Education*. Springer, 157–171.
- [47] Claire Hughes and Sue Leekam. 2004. What are the links between theory of mind and social relations? Review, reflections and new directions for studies of typical and atypical development. *Social Development* 13, 4 (2004), 590–619. <https://doi.org/10.1111/j.1467-9507.2004.00285.x>
- [48] C J Hutto and E E Gilbert. 2014. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014* (2014). <http://sentic.net/>
- [49] Kokil Jaidka, Sharath Chandra Guntuku, Anneke Buffone, H Andrew Schwartz, and Lyle H Ungar. 2018. Facebook vs. Twitter: Cross-platform differences in self-disclosure and trait prediction. In *Proceedings of the Twelfth International AAAI Conference on Web and Social Media*. 141–150.
- [50] Yuin Jeong, Younah Kang, and Juho Lee. 2019. Exploring effects of conversational fillers on user perception of conversational agents. *Conference on Human Factors in Computing Systems - Proceedings* (2019), 1–6. <https://doi.org/10.1145/3290607.3312913>
- [51] Da-jung Kim and Youn-kyung Lim. 2019. Co-Performing Agent: Design for Building User-Agent Partnership in Learning and Adaptive Services. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*. ACM Press, New York, New York, USA, 1–14. <https://doi.org/10.1145/3290605.3300714>
- [52] Kyung-Joong Kim and Hod Lipson. 2009. Towards a simple robotic theory of mind. (2009), 131. <https://doi.org/10.1145/1865909.1865937>
- [53] Soomin Kim, Jinsu Eun, Changhoon Oh, Bongwon Suh, and Joonhwan Lee. 2020. Bot in the Bunch: Facilitating Group Chat Discussion by Improving Efficiency and Participation with a Chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–13. <https://doi.org/10.1145/3313831.3376785>
- [54] Anastasia Kuzminykh, Jenny Sun, Nivetha Govindaraju, Jeff Avery, and Edward Lank. 2020. Genie in the Bottle: Anthropomorphized Perceptions of Conversational Agents. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–13. <https://doi.org/10.1145/3313831.3376665>
- [55] Sunok Lee, Sungbae Kim, and Sangsu Lee. 2019. "What does your Agent look like?". In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–6. <https://doi.org/10.1145/3290607.3312796>
- [56] Sangwon Lee, Naeun Lee, and Young June Sah. 2020. Perceiving a Mind in a Chatbot: Effect of Mind Perception and Social Cues on Co-presence, Closeness, and Intention to Use. *International Journal of Human-Computer Interaction* 36, 10 (2020), 930–940. <https://doi.org/10.1080/10447318.2019.1699748>
- [57] Séverin Lemaignan and Pierre Dillenbourg. 2015. Mutual modelling in robotics: Inspirations for the next steps. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 303–310.
- [58] Q. Vera Liao, Werner Geyer, Muhammed Mas-ud Hussain, Praveen Chandar, Matthew Davis, Yasaman Khazaeni, Marco Patricio Crasso, Dakuo Wang, Michael Muller, and N. Sadat Shami. 2018. All Work and No Play? Conversations with a Question-and-Answer Chatbot in the Wild. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, Vol. 8. ACM Press, New York, New York, USA, 1–13. <https://doi.org/10.1145/3173574.3173577>
- [59] Shuhong Lin, Boaz Keysar, and Nicholas Epley. 2010. Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention. *Journal of Experimental Social Psychology* 46, 3 (2010), 551–556. <https://doi.org/10.1016/j.jesp.2009.12.019>
- [60] Catherine L Lortie and Matthieu J Guittin. 2011. Judgment of the humanness of an interlocutor is in the eye of the beholder. *PLoS One* 6, 9 (2011), e25085.
- [61] Ewa Luger and Abigail Sellen. 2016. "Like having a really bad PA": the gulf between user expectation and experience of conversational agents. *CHI '16 Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016), 5286–5297. <https://doi.org/10.1145/2858036.2858288>
- [62] François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research* 30 (2007), 457–500.
- [63] Douglas R McCallum and James L Peterson. 1982. Computer-based readability indexes. In *Proceedings of the ACM '82 Conference*. 44–48.
- [64] Patrick E McKnight and Julius Najab. 2010. Kruskal-wallis test. *The corsini encyclopedia of psychology* (2010), 1–1.
- [65] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Neural Information Processing Systems (NIPS)*. 3111–3119.
- [66] Masahiro Mori, Karl F MacDorman, and Norri Kageki. 2012. The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine* 19, 2 (2012), 98–100.
- [67] Kellie Morrissey and Jurek Kirakowski. 2013. 'Realness' in chatbots: Establishing quantifiable criteria. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 8007 LNCS, PART 4 (2013), 87–96. [https://doi.org/10.1007/978-3-642-39330-3\\_10](https://doi.org/10.1007/978-3-642-39330-3_10)
- [68] Nora A. Murphy. 2007. Appearing smart: The impression management of intelligence, person perception accuracy, and behavior in social interaction. *Personality and Social Psychology Bulletin* 33, 3 (2007), 325–339. <https://doi.org/10.1177/0146167206294871>
- [69] Clifford Nass and Youngme Moon. 2000. Machines and mindlessness: Social responses to computers. *Journal of Social Issues* 1, 56 (2000), 81–103.
- [70] Clifford Nass, Jonathan Steuer, and Ellen Tauber. 1994. Computers are Social Actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 1994)*. ACM. <https://doi.org/10.1109/VISM.2014.7136659>
- [71] Oda Elise Nordberg, Jo Dugstad Wake, Emilie Sektan Nordby, Eivind Flobak, Tine Nordgreen, Suresh Kumar Mukhiya, and Frode Guribye. 2020. Designing Chatbots for Guiding Online Peer Support Conversations for Adults with ADHD. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 11970 LNCS, November (2020), 113–126. [https://doi.org/10.1007/978-3-030-39540-7\\_18](https://doi.org/10.1007/978-3-030-39540-7_18)
- [72] Nicole Novielli, Fiorella de Rosi, and Irene Mazzotta. 2010. User attitude towards an embodied conversational agent: Effects of the interaction mode. *Journal of Pragmatics* 42, 9 (2010), 2385–2397. <https://doi.org/10.1016/j.pragma.2009.12.016>
- [73] Catherine Pelachaud and Isabella Poggi. 2002. Subtleties of facial expressions in embodied agents. *The Journal of Visualization and Computer Animation* 13, 5 (2002), 301–312.
- [74] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.
- [75] Christopher Peters. 2005. Foundations of an agent theory of mind model for conversation initiation in virtual environments. *Virtual Social Agents* (2005), 163.
- [76] Matthew D. Pickard, Judee K. Burgoon, and Douglas C. Derrick. 2014. Toward an Objective Linguistic-Based Measure of Perceived Embodied Conversational Agent Power and Likeability. *International Journal of Human-Computer Interaction* 30, 6 (2014), 495–516. <https://doi.org/10.1080/10447318.2014.888504>
- [77] Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 186–195.
- [78] David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences* 1, 4 (1978), 515–526.
- [79] David V. Pynadath and Stacy C. Marsella. 2005. PsychSim: Modeling theory of mind with decision-theoretic agents. *IJCAI International Joint Conference on Artificial Intelligence* (2005), 1181–1186.
- [80] Stephen Reysen. 2005. Construction of a new scale: The Reysen likability scale. *Social Behavior and Personality: an international journal* 33, 2 (2005), 201–208.
- [81] Tina L Robbins and Angelo S DeNisi. 1994. A closer look at interpersonal affect as a distinct influence on cognitive processing in performance evaluations. *Journal of Applied Psychology* 79, 3 (1994), 341.
- [82] Sherry Ruan, Liwei Jiang, Justin Xu, Bryce Joe-Kun Tham, Zhengneng Qiu, Yeshuang Zhu, Elizabeth L Murnane, Emma Brunskill, and James A Landay. 2019. Quizbot: A dialogue-based adaptive learning system for factual knowledge. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.
- [83] Koustuv Saha, Manikanta D Reddy, Stephen Mattingly, Edward Moskal, Anusha Sirigiri, and Munmun De Choudhury. 2019. Libra: On linkedin based role ambiguity and its relationship with wellbeing and job performance. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–30.
- [84] Koustuv Saha and Amit Sharma. 2020. Causal Factors of Effective Psychosocial Outcomes in Online Mental Health Communities. In *ICWSM*.
- [85] Koustuv Saha, Ingmar Weber, and Munmun De Choudhury. 2018. A Social Media Based Examination of the Effects of Counseling Recommendations After Student Deaths on College Campuses. In *ICWSM*.

- [86] Joseph Seering, Juan Pablo Flores, Saiph Savage, and Jessica Hammer. 2018. The social roles of bots: Situating bots in discussions in online communities. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018). <https://doi.org/10.1145/3274426>
- [87] Joseph Seering, Michal Luria, Geoff Kaufman, and Jessica Hammer. 2019. Beyond dyadic interactions: Considering chatbots as community members. In *Conference on Human Factors in Computing Systems - Proceedings*. <https://doi.org/10.1145/3290605.3300680>
- [88] Joseph Seering, Michal Luria, Connie Ye, Geoff Kaufman, and Jessica Hammer. 2020. It Takes a Village: Integrating an Adaptive Chatbot into an Online Gaming Community. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–13. <https://doi.org/10.1145/3313831.3376708>
- [89] James Simpson. 2020. Are CUIs Just GUIs with Speech Bubbles?. In *Proceedings of the 2nd Conference on Conversational User Interfaces*. 1–3.
- [90] Marcin Skowron, Stefan Rank, Mathias Theunis, and Julian Sienkiewicz. 2011. The good, the bad and the neutral: affective profile in dialog system-user communication. In *International Conference on Affective Computing and Intelligent Interaction*. Springer, 337–346.
- [91] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.
- [92] Qiaosi Wang, Shan Jing, Ida Camacho, David Joyner, and Ashok Goel. 2020. Jill Watson SA: Design and Evaluation of a Virtual Agent to Build Communities Among Online Learners. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–8.
- [93] Qiaosi Wang, Shan Jing, David Joyner, Lauren Wilcox, Hong Li, Thomas Plötz, and Betsy Disalvo. 2020. Sensing Affect to Empower Students: Learner Perspectives on Affect-Sensitive Technology in Large Educational Contexts. In *Proceedings of the Seventh ACM Conference on Learning@ Scale*. 63–76.
- [94] Henry M Wellman. 1992. *The child's theory of mind*. The MIT Press.
- [95] Rainer Winkler, Sebastian Hobert, Antti Salovaara, Matthias Söllner, and Jan Marco Leimeister. 2020. Sara, the Lecturer: Improving Learning in Online Education with a Scaffolding-Based Conversational Agent. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–14. <https://doi.org/10.1145/3313831.3376781>
- [96] Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, and Rama Akkiraju. 2017. A new chatbot for customer service on social media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 3506–3510.
- [97] Xi Yang, Marco Aurisicchio, and Weston Baxter. 2019. Understanding Affective Experiences with Conversational Agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*. ACM Press, New York, New York, USA, 1–12. <https://doi.org/10.1145/3290605.3300772>
- [98] Jennifer Zamora. 2017. I'm sorry, dave, i'm afraid i can't do that: Chatbot perception and expectations. In *Proceedings of the 5th International Conference on Human Agent Interaction*. 253–260.
- [99] Justine Zhang, Jonathan P Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Nithum Thain, and Dario Taraborelli. 2018. Conversations gone awry: Detecting early signs of conversational failure. *arXiv preprint arXiv:1805.05345* (2018).

## A APPENDIX

This material (Figure 3) presents the bi-weekly perception survey students filled out. It was adapted from [Bartneck et al.](#) for measuring human-robot interaction. We particularly selected the categories of anthropomorphism, intelligence, and likeability in our setting of student perceptions about JW.

The following questions will give you a spectrum from one quality to the other on a scale of 1 to 5, such as from "Unkind"(1) to "Kind"(5). Please rate your perception of JW along each of these spectrums:

Fake 1 2 3 4 5 Natural	✓ [ Select ] 1 2 3 4 5
Unintelligent 1 2 3 4 5 Intelligent	[ Select ]
Unkind 1 2 3 4 5 Kind	[ Select ]
Foolish 1 2 3 4 5 Sensible	[ Select ]
Artificial 1 2 3 4 5 Lifelike	[ Select ]
Dislike 1 2 3 4 5 Like	[ Select ]
Awful 1 2 3 4 5 Nice	[ Select ]
Ignorant 1 2 3 4 5 Knowledgeable	[ Select ]
Machinelike 1 2 3 4 5 Humanlike	[ Select ]
Responding rigidly 1 2 3 4 5 Responding elegantly	[ Select ]
Unfriendly 1 2 3 4 5 Friendly	[ Select ]
Irresponsible 1 2 3 4 5 Responsible	[ Select ]
Unpleasant 1 2 3 4 5 Pleasant	[ Select ]
Incompetent 1 2 3 4 5 Competent	[ Select ]
Unconscious 1 2 3 4 5 Conscious	[ Select ]

Figure 3: Anthropomorphism items are marked with green boxes, intelligence items were marked with orange boxes, and likeability items were marked with blue boxes.