Does Chatbot Language Formality Affect Users' Self-Disclosure?

Samuel Rhys Cox
samcox@comp.nus.edu.sg
National University of Singapore
Singapore

Wei Tsang Ooi
ooiwt@comp.nus.edu.sg
National University of Singapore
Singapore

ABSTRACT

Chatbots are increasingly used to replace human interviewers and survey forms for soliciting information from users. This paper presents two studies that investigate how the formality of a chatbot's conversational style can affect the likelihood of users engaging with and disclosing sensitive information to a chatbot. In our first study, we show that the domain and sensitivity of the information being requested impact users' preferred conversational style. Specifically, when users were asked to disclose sensitive health information, they perceived a formal style as more competent and appropriate. In our second study, we investigate the health domain further by analysing the quality of user utterances as users talk to a chatbot about their dental flossing. We found that users who do not floss every day gave higher quality responses when talking to a formal chatbot. These findings can help designers choose a chatbot's language formality for their given use case.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; • **Security and privacy** → **Social aspects of security and privacy**.

KEYWORDS

Chatbots, Conversational Agents, Conversational Style, Language Formality, Information Sensitivity, Self-disclosure

ACM Reference Format:
Samuel Rhys Cox and Wei Tsang Ooi. 2022. Does Chatbot Language Formality Affect Users' Self-Disclosure?. In *4th Conference on Conversational User Interfaces (CUI 2022), July 26–28, 2022, Glasgow, United Kingdom.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3543829.3543831

1 INTRODUCTION

Chatbots can be used to replace the function of surveys and interviews as a form of data collection [43, 87]. In doing so, chatbots harness some of the advantages of surveys such as scalability and low-cost; as well as some of the advantages of human interviews such as dialogue based interactions, and human-like characteristics.

However, user attitudes towards chatbots may be sculpted by their expectations of the most appropriate conversational style, and using the wrong conversational style for a given situation can lead to user frustration [15, 25]. To overcome this issue, previous work has built upon the Computers Are Social Actors (CASA) paradigm [57], which asserts that people apply social norms from human interactions when interacting with computer agents. This has led to the design of conversational agents that exhibit human-like characteristics that users perceive more positively. For example, previous work has investigated the use of different conversational styles when requesting information [5, 6, 37, 82], and the use of self-disclosing chatbots to encourage reciprocated user self-disclosure [1, 47, 56].

Despite this, the effect of a chatbot's language formality on the quality of a user's self-disclosure has not been investigated. Previous work has already produced some conflicting results regarding the formality of a chatbot's utterances. For example, participants criticised a HIV chatbot for being overly formal and dissimilar to how people normally talk when using chat applications [81]. In contrast, users noted that the tone of a chatbot to perform financial services was too informal and not serious enough given the task at hand [24]. In studies more analogous to self-disclosure, it has been found that chatbots exhibiting a casual conversational style can lead to higher quality collection of multiple choice responses compared to standardised surveys [11, 43, 83]. Yet, these studies did not explicitly compare different levels of language formality, and instead compared the script from a standardised (non-conversational) survey against a casual (conversational) script. Thus, these studies motivate an interesting issue, and inspire the **high-level research question** for our paper:

- What is the effect of a chatbot's language formality on a user's self-disclosure?

To evaluate this question, we carried out two experiments on Amazon Mechanical Turk (AMT). Both studies investigated the impact of using either a casual or (comparatively) formal conversational style, and we used an objective measure of language formality [35] to design chatbot scripts. Both studies are motivated similarly at a high-level to investigate the effect of a chatbot's language formality on self-disclosure; with Study 1 investigating *likelihood* to disclose information, and Study 2 investigating the *quality* of user utterances (and thereby the quality of self-disclosure to the chatbot).

Specifically, in Study 1 (N = 187) we measured user perceptions and the likelihood of information disclosure while varying the information requested by the chatbot by its domain (health vs non-health information) and information sensitivity (sensitive vs non-sensitive information). We used empirically validated levels of information sensitivity to differentiate our levels [54], and (as we did not want to collect user's sensitive information) we conducted a hypothetical scenario-driven experiment where participants indicated their likelihood to disclose information to the chatbot (similarly to [73]).

CUI 2022, July 26–28, 2022, Glasgow, United Kingdom

ACM ISBN 978-1-4503-9739-1/22/07…$15.00
https://doi.org/10.1145/3543829.3543831

The findings of Study 1 indicated the benefits of varying a chatbot's language formality when requesting health information. However, as these findings were hypothetical, it motivated us to conduct Study 2 ($N = 156$) where we analysed the quality of user utterances when people discuss health behaviours with a chatbot.

From our two user studies, we found evidence that chatbots related to the health domain could benefit by adopting a more formal conversational style. Specifically, Study 1 found that when a chatbot elicits sensitive health information, users perceive a formal conversational style as more competent and appropriate than a comparatively casual style. Study 2 found that a chatbot that adopts a formal conversational style elicits higher quality user utterances compared to a comparatively casual conversational style.

2 RELATED WORK

In this section we discuss how higher quality responses can be elicited from users of chatbots, and how the language used by a chatbot can affect user perceptions and engagement.

2.1 Designing Chatbots to Act Like People

Prior studies have created chatbots that replicate conversational styles people perceive positively. This has in part been influenced by the CASA paradigm [57], which states that people apply social norms from human interactions when interacting with computer agents. On from this, studies have focused on producing human-like chatbots that emulate human responses and contextual awareness, such as by creating chatbots using empathic language [32, 49, 51, 67] or politeness strategies [20, 22, 72]. Drawing on CASA, we can develop chatbots that incorporate recommendations for human-to-human communication, but it is uncertain whether such recommendations apply to chatbots given differences in perceived social roles and user expectations of chatbots [42]. Conflicting findings for [11, 43, 81, 83] and against [24, 40] using casual chatbot conversational styles thereby motivate us to compare the impact of a chatbot's language formality on user self-disclosure.

2.2 Improving Self-Disclosure to Chatbots

Now, we will discuss work that has manipulated chatbot features to increase user self-disclosure (the voluntary sharing of information from one person to another, such as feelings, opinions, or personal information [63]). People often make decisions to disclose information based on cognitive heuristics: general rules of thumb users apply to different situations [27, 73, 74, 80]. For example, Sundar et al. found the effect of the machine heuristic applies to disclosure to a chatbot, specifically: people were more likely to disclose to a chatbot than to a human because they believe a machine handles information more securely [73].

Previous studies have investigated the effects of changing a chatbot's traits such as personality [88], levels of mutual disclosure [1, 47, 56, 70], and use of small talk [5, 6, 37, 82] on users' self-disclosure. Zhou et al. developed chatbots for job interviews that had either a warm, cheerful personality or a serious, assertive personality, and found that high-stakes job interviewees were more willing to confide in a chatbot that used a serious and assertive personality [88]. Adam et al. found people disclose more to a chatbot when it requests information conversationally one-by-one rather than all at once (similarly to a standard survey) [1]. Lee et al. found people disclosed more to a mental health chatbot that disclosed information about itself [47]. Brickmore et al. found that conversational agents engaging in social dialogue (such as small talk) lead to increased levels of rapport and self-disclosure [5, 6]. Xiao et al. used Gricean maxims [31] (quantity, informativeness, relevancy, manner) to measure the quality of user utterances [86, 87], and found people gave higher quality utterances to a chatbot using active listening techniques compared to a chatbot giving generic responses [86].

On from this, our user studies measured the effect of a chatbot's language formality both on a user's likelihood to disclose information (similarly to Sundar et al. [73]), and quality of user utterances (similarly to Xiao et al. [86, 87]).

2.3 Changing Chatbot Conversational Style

As we want to investigate the effect of varying a chatbot's conversational style (i.e., the formality of a chatbot's utterances), we will provide an overview of related literature.

Previous work has investigated the effect of chatbots incorporating different conversational styles [11, 25, 43, 50, 83]. For example, Elholz et al. compared user perceptions of two theatre ticket booking chatbots: one using a modern, and one using a Shakespearean conversational style [25]. They found that while the former was more user-friendly, the latter was more entertaining to users. Kim et al. compared the effect of a chatbot using a casual conversational script (adopting conversational language, emojis and abbreviations) against a chatbot using a standardised survey script, and found that a casual conversational style leads to collection of higher quality multiple choice survey responses [43]. However, studies using casual conversational styles [11, 43, 83] used baseline conditions of language from traditional surveys (using more terse, non-conversational language) rather than a formal but conversational style, and did not investigate the effect of asking for personally sensitive information.

We therefore conducted user studies where a chatbot elicited information using comparatively casual and formal conversational styles to investigate the effect of a chatbot's language formality on user perceptions and self-disclosure.

3 STUDY 1: THE EFFECT OF CONVERSATIONAL STYLE ON DISCLOSURE

In this study, we investigate how the conversational style of a chatbot and the sensitivity of information requested by the chatbot affect user perceptions and likelihood to disclose information. This gave us the research questions of:

- **RQ1:** How does the domain and sensitivity of information requested by a chatbot influence a user's likelihood to disclose information?
- **RQ2:** How does a chatbot's conversational style (formal vs casual) influence a user's likelihood to disclose information?

For this, we conducted a between-subjects, scenario-based experiment on AMT where participants were asked to imagine they had been talking to a chatbot to help them discover life insurance

products. Life insurance was chosen as the chatbot's use case as this allowed the chatbot to request information of varying sensitivity, and ensured the scenario was relatable given the prevalence of financial assistance chatbots.

3.1 Experiment Conditions

Participants were shown a screenshot of a conversation between a human and a chatbot. Chatbot scenarios shown to participants were 2 × 3 factorial with the following factors:

- *Chatbot Conversational Style*: casual, formal;
- *Domain Sensitivity (of information requested)*: Low-sensitivity non-health, High-sensitivity non-health, High-sensitivity health.

3.1.1 Chatbot Conversational Style. To objectively measure language formality, we use a formality measure taken from linguistics literature [35], which has been used previously to differentiate between genres of texts [58, 76], and detect clickbait news articles [7]. This metric (known as the F-score) calculates language formality by using the part of speech of language in the following formula:

> F = (noun frequency + adjective freq. + preposition freq. + article freq. − pronoun freq. − verb freq. − adverb freq. − interjection freq. + 100)/2

This formula measures the "deep formality" (akin to the specificity) of a text. From this definition, a casual style uses more deictic terms whose understanding is dependant upon context (such as "this", "later"), whereas a formal style uses explicit terms such as nouns, prepositions, and articles. A higher F-score denotes more formal language, while a lower F-score is more casual. Within Heylighen's [35] study (for the English language) phone conversations had the lowest F-score (scoring 36), while informational writing had the highest F-score (scoring 61). Although the F-score has limitations such as not accounting for differences in punctuation, and scoring all words as ±1 (while slang could be perceived as more casual and technical language more formal), it acts as a proxy of formality to differentiate our levels of conversational style.

This definition of formality should not be confused with references to casual language in other literature, which are more analogous to using politeness strategies, empathic communication, emojis, or slang [24, 40, 43, 77]. As we are not investigating the effectiveness of these, we scripted chatbot dialogue to be similar (in terms of meaning and politeness) between casual and formal conditions. See Figure 1 for the chat sessions shown to participants in the casual and formal conditions, along with each condition's corresponding F-score.

3.1.2 Domain Sensitivity of Information Requested. We compared the influence of low and high sensitivity information, and health and non-health information. Doing so allowed us to investigate whether people dislike a request for information due to information requested being health sensitive or information sensitive. To choose experiment conditions, we drew from Markos et al.'s survey of US consumers' perceived sensitivity of their personal information [54]. Specifically, we chose three pieces of information which affect the price of life insurance policies as our experiment conditions, giving us conditions where the chatbot asks the user for their:

- **income level** (low-sensitivity non-health [54])
- **credit score** (high-sensitivity non-health [54]).
- **medical history** (high-sensitivity health [54]).

3.2 Design Considerations

To avoid racial or gender mirroring effects [48], we used a robot icon to represent the chatbot in the mock-up scenarios (see Figure 1). Dialogue in all conditions was the same except for the language formality of chatbot responses, and the information requested by the chatbot in the final message. Human responses are identical between conditions, and chatbot scripts deploy the same politeness strategies [8, 20] and follow the same conversation structure. First, the chatbot greets the user, states its purpose, and offers assistance. Next, the chatbot asks for permission to request information, before finally requesting specific information.

3.3 Participants

We recruited participants from AMT, and used selection criteria to ensure reliable results (US based, >98% approval rate, >5,000 completed HITs). Participants were paid USD$0.60 and the survey took a mean time of 4.22 minutes to complete. 264 participants attempted the survey, and 187 (mean age 37.4; 81 female) passed screening tasks. Participants were randomly assigned a condition with 95 in casual and 92 in formal.

3.4 Procedure

Participants completed the following procedure:

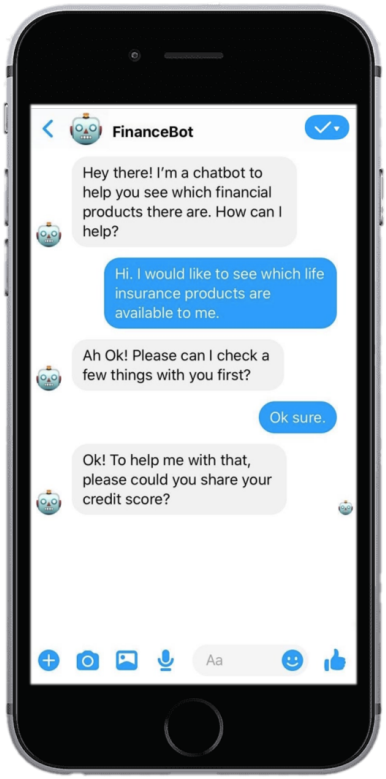
(1) Introduction and consent form.
(2) Pass word associativity test [12] screening for English language skills (similarly to previous crowdsourcing studies [12, 18]).
(3) Answer demographics (age, gender, education level, prior experience with chatbots) and one instructional manipulation check (from Oppenheimer et al. [59]), followed by more detailed instructions.
(4) Each participant sees one screenshot of a conversation between a chatbot and a user seeking information about financial products (see Figure 1 for screenshots). Participants were told "*Please imagine you have been talking to the chatbot in the conversation below*", and evaluated the chat sessions using measures described in Section 3.5.
(5) Post-test questions (see Section 3.6).

3.5 Measures

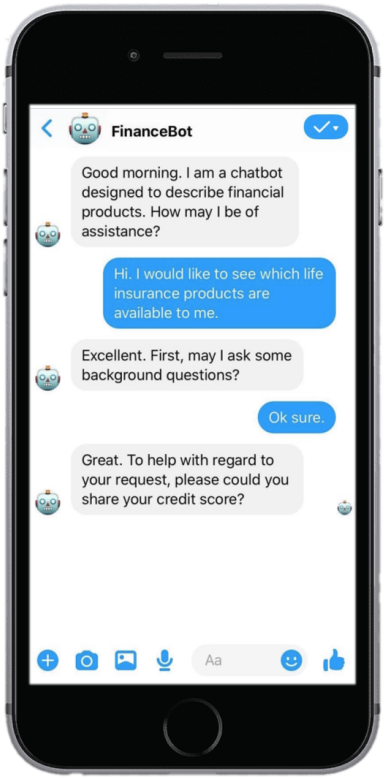
Participants evaluated the scenario assigned to them using measures on 7-point Likert scales (1 = Strongly Disagree to 7 = Strongly Agree), and were asked "*Do you personally agree or disagree that...*" for the following:

- **Likelihood to disclose**: "I am likely to disclose my ["income level"/"credit score"/"medical history"] to the chatbot" [54, 73]
- **Enjoyment**: "I would enjoy talking to the chatbot" [11]
- **Warmth**: "This chatbot was warm", "This chatbot was good-natured" [19] (M = 4.85, SD = 1.31, α = 0.865)[1]

[1]Here α denotes the Cronbach's alpha reliability score (the extent to which a measure is consistent). Values above 0.8 are considered good for empirical studies [46].



(a) Casual chat condition: with F-score of 42 (similar to F-score of normal spoken English conversation [35])



(b) Formal chat condition: example with F-score of 50.5 (similar to F-score of prepared English speech [35])

Figure 1: The chatbot conversation for both casual and formal conversational styles when asking for the user's credit score.

- **Competency**: "This chatbot was competent", "This chatbot was capable" [19] (M = 5.26, SD = 1.16, α = 0.965)
- **Appropriateness of tone**: "The chatbot used an appropriate tone" [28]

Likelihood to disclose was used as a direct measure for RQ1 and RQ2. Other subjective measures were collected to gain additional insights, but do not explicitly measure self-disclosure. Similarly to previous chatbot studies [42, 52], warmth and competency measures are taken from the Stereotype Content Model (SCM) [19], that posits that human stereotypes are formed along two dimensions of warmth and competence. Specifically, warmth is the intention of someone to help you, and competence is their ability to act on these intentions. Groups perceived as both warm and competent inspire desire to assist them. For example, physicians are perceived as both warm and competent.

Additionally, free-text feedback was collected asking participants *why* they liked or disliked chatbot conversations ("What do you like or dislike about the conversation with the chatbot?"). See Section 4.1 for qualitative findings.

3.6 Post Interaction Survey

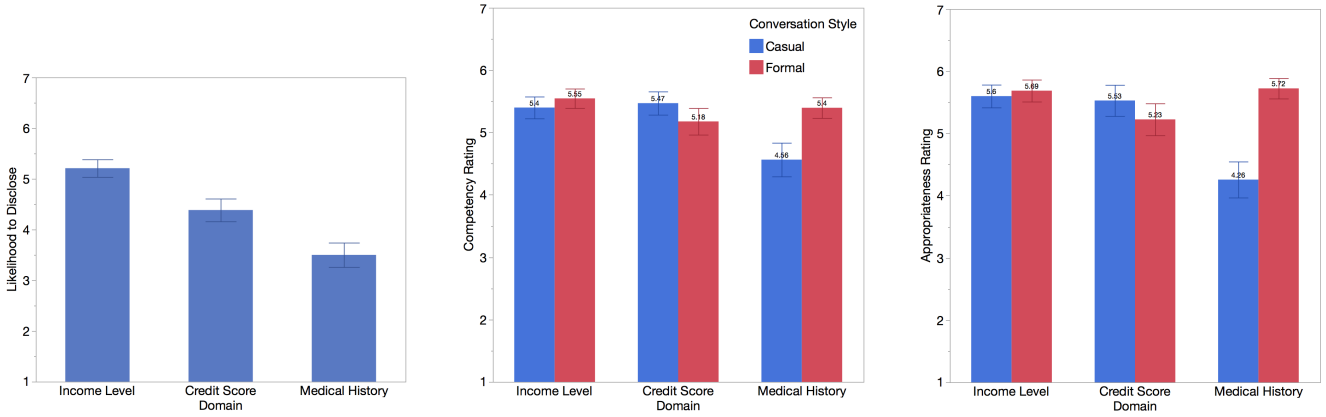
On experiment completion, participants answered a post-interaction survey measuring levels of privacy concerns, belief in robotic intelligence, and belief in robotic feelings (measured on 7-point Likert scales - 1 = Strongly Disagree to 7 = Strongly Agree).

Levels of privacy concerns were measured using 4 items from Dinev and Hart [21]: "*I am concerned that the information I submit on the Internet could be misused.*", "*I am concerned about submitting information on the Internet, because of what others might do with it.*", "*I am concerned about submitting information on the Internet, because it could be used in a way I did not foresee.*", and "*I am concerned that a person can find private information about me on the Internet.*" (M = 5.01, SD = 1.60, α = 0.962).

To measure belief in robotic intelligence, participants responded to "*I believe robots can be intelligent*" and to measure belief in robotic feelings participants responded to "*I believe robots can have real feelings*" (both from Liu et al. [51]).

4 STUDY 1: RESULTS AND FINDINGS

We fit a linear model on each dependent variable with conversational style and domain as the fixed effects, and performed post-hoc Tukey's HSD to identify specific differences. See Table 1 results



(a) Participants were most likely to disclose income level and least likely to disclose medical history

(b) When asked for their medical history, participants perceived a chatbot as more competent if it used a formal conversational style

(c) When asked for their medical history, participants perceived a chatbot as more appropriate if it used a formal conversational style

Figure 2: Key figures for Study 1. Graphs 2b and 2c show significant differences between casual and formal conversational styles when medical history was requested.

	Income Level		Credit Score		Medical History	
	Casual	Formal	Casual	Formal	Casual	Formal
Likelihood to Disclose	5.03 (1.58)	5.34 (1.13)	4.41 (1.88)	4.35 (1.76)	3.51 (1.95)	3.48 (1.80)
Enjoyment	4.33 (1.40)	4.56 (1.44)	4.32 (1.47)	4.13 (1.43)	3.94 (1.70)	4.10 (1.72)
Warmth	5.05 (1.23)	4.84 (1.09)	4.92 (1.17)	4.60 (1.40)	4.85 (1.37)	4.83 (1.12)
Competence	5.40 (0.96)	5.55 (0.88)	5.47 (1.09)	5.18 (1.19)	**4.56 (1.50)**	**5.40 (0.89)** **
Appropriateness of Tone	5.60 (1.00)	5.69 (1.00)	5.53 (1.46)	5.23 (1.43)	**4.26 (1.61)**	**5.72 (0.88)** **

Table 1: Outcome measures for Study 1 split by both Domain and Conversational Style (values shown as "Mean (s.d.)"). Significant differences shown in bold ($p < .0001$).

summary by domain and conversational style. We will now discuss specific findings and their implications, before discussing qualitative findings in Section 4.1.

Similarly to previous findings within privacy literature [54], the domain of information requested influenced the likelihood of information disclosure (see Figure 2a). Participants were most likely to disclose income level (M = 5.21), less likely to disclose credit score (M = 4.39), and least likely to disclose medical history (M = 3.50) ($p < .0001$). While Markos et al. found participants equally likely to disclose "medical history" and "credit rating", we found participants were less likely to disclose their "medical history". This difference may be due to how concretely people can perceive the information being requested (i.e., medical history may appear more open-ended), or due to people questioning the necessity of the information being requested (similar to findings from Heckler et al. where they noted the perceived sensitivity of health information was relative to the domain from which it is being requested [34]).

While conversational style had no direct impact on likelihood to disclose, conversational style did influence user perceptions. There was a significant effect of (conversational style against domain) on perceived competence and appropriateness. Post-hoc tests found that when being asked for their medical history, participants found Formal more competent ($p < .0001$) and appropriate ($p < .0001$) compared to Casual (see Figures 2b and 2c). There were moderate significant effects for conversational style ($p = 0.0228$) and domain ($p = 0.0381$) on appropriateness. The difference for appropriateness between Casual (M = 5.14) and Formal (M = 5.54) may seem to suggest that a formal conversational style is always preferred, but the difference can be accounted for by the difference in medical history (see Figure 2c).

Participants with lower levels of privacy concerns found the chatbot more enjoyable ($p < .0001$), warm ($p = 0.0045$), competent ($p = 0.0034$) and appropriate ($p = 0.0008$) and were more likely to disclose to the chatbot ($p < .0001$) than those with high

Table 2: Frequency count of categories within user feedback. Casual was more often described as human-like or unprofessional, while Formal as polite or professional.

Category	Total	Casual	Formal
Human-like	41	25	16
Uncanny	19	9	10
Friendly	56	31	25
Unfriendly	6	1	5
Professional	41	9	32
Unprofessional	12	12	0
Polite	26	5	21
Helpful	15	7	8

levels of privacy concerns. Participants who believe in robotic intelligence found the chatbot more enjoyable ($p < .0001$), warm ($p < .0001$), competent ($p < .0001$), and appropriate ($p = 0.0058$) than participants who do not believe in robotic intelligence. Similarly, participants who believe in robotic feelings found the chatbot more enjoyable ($p = 0.0004$) and were more likely to disclose to the chatbot ($p = 0.0033$) than participants who do not believe in robotic feelings.

Overall, these findings suggest that the likelihood to disclose is affected by the perceived sensitivity of the information being requested rather than the conversational style of the agent requesting the information. However, the finding that people perceived a formal conversational style as more competent and appropriate when sensitive health information was requested suggests that there may be more imperceptible effects taking place.

4.1 Qualitative findings from user feedback

To investigate further why people perceive conversational styles differently we coded user feedback. All participant feedback was open-coded by a member of the research team blind to the experiment condition of the feedback. Due to varying levels of detail provided in the feedback, each response could be coded into one, multiple, or no categories. This gave an initial set of 29 categories, which were discussed and consolidated into 8 categories. A second researcher from outside the project then blind labelled the feedback using the 8 consolidated categories. After this, 187 of the labels were identical between the two labellers, while 29 labels conflicted (i.e., an alternative or additional category was labelled for a given piece of feedback). These conflicting labels were discussed and finalised between the two labellers, giving us the final category count found in Table 2.

From this coding, we found that Casual was more often described as human-like:

> "*I like how the chatbot uses informal syntax to make it feel more human-like.*" -P125-Casual(Income Level)

Several Casual participants enjoyed the friendly nature of conversation, but expressed concern for level of professionalism:

> "*I like that the chat bot seems friendly and personable but I would not want that to go too far and become unprofessional.*" -P222-Casual(Income Level)

This perceived unprofessionalism was reinforced by several participants who said the casual tone was not domain appropriate, or would affect their likelihood to disclose:

> "*I dislike that the language style of the chatbot is not professional and that makes me cautious of the type of information that I have to provide.*" -P82-Casual(Credit Score)
>
> "*It's too jovial and colloquial for such a topic. Friendly is fine if I'm buying clothes but not when discussing finances.*" -P2-Casual(Income Level)

In contrast, Formal was more often described as professional or polite:

> "*I like how the bot was talking in a professional manner. This makes me trust it more and not think it's a terrible idea to use chat bots.*" -P96-Formal(Credit Score)
>
> "[...] *It is respectful in asking for certain information from you that is otherwise private.*" -P122-Formal(Credit Score)

However, participants also noted Formal lacked empathy:

> "*It's not very reassuring or comforting*" -P152-Formal(Medical History)
>
> "*It seems very impersonal, which may come across as cold, but that also means it doesn't have any particular bias*" -P10-Formal(Medical History)

Interestingly, Formal(*Medical History*) feedback suggests while participants may find tone more appropriate, they would not be willing to disclose their personal information in an online chat setting:

> "*The language sounded patient and warm and the chat bot made me feel comfortable. But, I would still feel uncomfortable sharing my medical information online in a chat room.*" -P19-Formal(Medical History)

In contrast, participants described Casual(*Medical History*) not meeting their expected level of appropriateness:

> "*The tone of the chatbot was a little too informal for the topic it was discussing. Discussing personal finances should have a more formal tone.*" -P82-Casual(Medical History)

Overall, the feedback indicates the importance of a conversational style matching a user's expectations. A conversational style that matches user expectations can lead to greater feelings of trust, while a mismatch leads to negative user reactions (such as Casual being perceived as unprofessional).

5 STUDY 2: THE EFFECT OF CONVERSATIONAL STYLE ON RESPONSE QUALITY

While Study 1 results did not find a difference in likelihood to disclose between conversational styles, we found that people perceived a formal conversational style as more appropriate and competent when being asked to disclose sensitive health information (i.e., "medical history").

However, as the information being requested in Study 1 was hypothetical, we next want to investigate the effect of a chatbot's

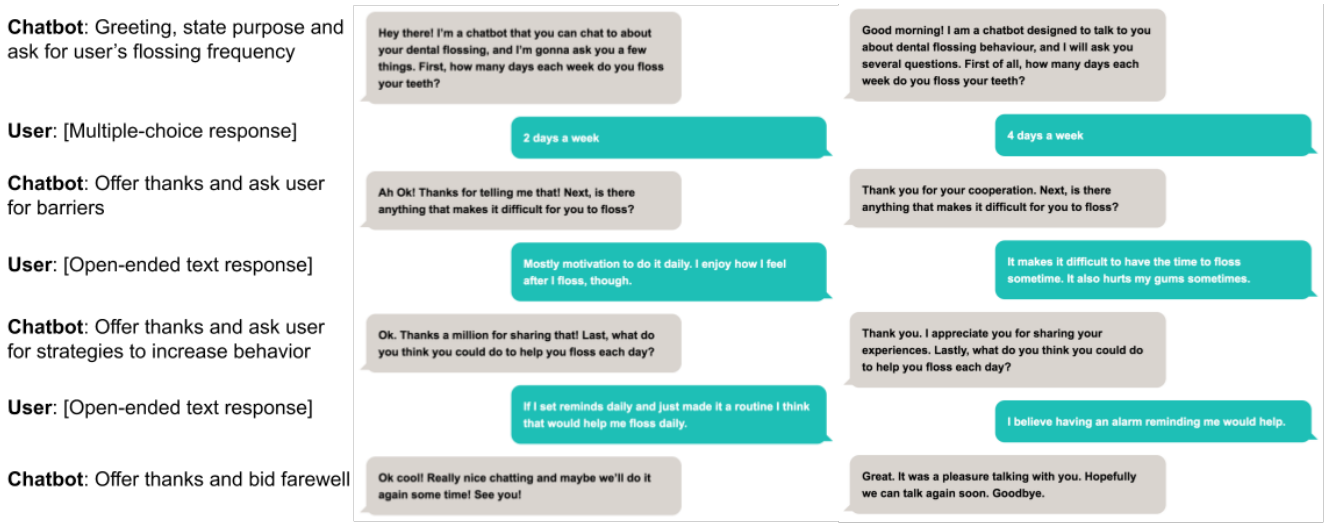


Figure 3: The Study 2 chatbot for both casual and formal conversational style conditions. The interface is as seen by participants with chatbot utterances in the grey speech bubbles and actual participant responses in green speech bubbles. The F-score for the casual condition was 40.5, and the F-score for the formal condition was 47.5.

language formality on the quality of user utterances. Therefore, we expanded on the Study 1 results for health information and ran a second study where participants discussed their health-related habits (specifically, their dental flossing) with a chatbot. We then analysed user utterances to investigate the effect of a chatbot's language formality on the quality of a user's responses. This gave us our third research question:

- **RQ3:** How does a chatbot's conversational style (formal vs casual) influence the quality of user utterances?

We chose not to request high sensitivity information (such as medical history), or to discuss health habits which are more likely to affect a participant's health (such as medication adherence) in order to lower privacy concerns and health implications of the study.

Accordingly, we chose dental flossing as the domain of our chatbot as we wanted to investigate a use case where a participant could talk to a chatbot about a healthy behaviour to which health experts recommend regular adherence (as is the case for flossing [71]). Additionally, people can have barriers to dental flossing [2, 9], which was key to our chatbot's script; people are benefitted by brief interventions [29] and daily diary keeping [75] - tasks which could be facilitated by chatbots; and flossing benefits one's health and is effective at reducing disease causing bacteria [17] and gum bleeding [30].

5.1 Experiment Conditions

We conducted a between-subjects experiment on AMT where participants talked to a chatbot using one of two conversational styles: casual or formal.

We want to control the level of formality of the responses given by the chatbot. To achieve this level of control, machine generated responses would not be appropriate. We therefore implemented a chatbot following rule-based logic and dialogue written by the research team. The chatbot was hosted within Qualtrics, and used HTML and JavaScript to emulate the look and feel of a chatbot. While intent classification could be used to recognise user intent and give adaptive chatbot responses, we chose to use pre-scripted responses so participants would see consistent responses between conditions and to reduce the potential for negative perceptions due to user intents being misclassified. Therefore, questions asked by the chatbot are written to avoid dependency to one another. Additionally, chatbot responses were given instantaneously independent of the length or complexity of the user's utterances.

Chatbot dialogue followed the same high-level structure between conditions (see Figure 3 for high-level chatbot utterances, conversational styles, and interface seen by participants). As part of this dialogue, the chatbot asks three questions. First is a multiple choice question to ask the user for their flossing frequency. Next, the chatbot asks two open-text questions asking the user to identify: (1) barriers to dental flossing, and (2) strategies for overcoming barriers to dental flossing. The phrasing of this second question depends on the flossing frequency of the user. Users who do not floss each day are asked "*what do you think you could do to help you floss each day?*" whereas those who floss every day are asked "*what do you do to help you floss each day?*". To ensure questions asked by the chatbot are interpreted identically across conditions, we used the same questions verbatim for each condition. The text preceding the question asked by the chatbot was varied in conversational style to be either casual or formal.

5.2 Participants

We recruited participants from AMT using the selection criteria described in Section 3.3. Participants were paid USD$0.85, and the survey took a mean time of 4.02 minutes to complete. 183 participants attempted the survey, and 156 (mean age 37.4; 58 female) passed the screening tasks of whom 44 used dental floss every day.

Participants were randomly assigned a condition with 80 in casual and 76 in formal.

5.3 Procedure

Participants completed the following procedure:

(1) Introduction and consent form.
(2) The same demographics and screening questions were used as those in Study 1.
(3) Remaining participants were shown more detailed task instructions, before conversing with the chatbot about their dental flossing, and rating their experience (see Section 5.4).
(4) Post-test questions (see Section 5.6).

5.4 Subjective Measures

Akin to Study 1, participants evaluated their chatbot interaction on 7-point Likert scales (1 = Strongly Disagree to 7 = Strongly Agree), and (while these measures are not explicitly linked to RQ3) they are intended to gain additional insights. Participants were asked "*Do you personally agree or disagree that...*" for the following:

- **Desire to continue**: "I would want to continue using the chatbot" [42]
- **Enjoyment**: "I enjoyed talking to the chatbot" [11]
- **Warmth**: "The chatbot was warm", "The chatbot was good-natured" [19] (M = 5.02, SD = 1.50, α = 0.894)
- **Competence**: "The chatbot was competent", "The chatbot was capable" [19] (M = 4.59, SD = 1.61, α = 0.930)
- **Appropriateness of tone**: "The tone of the chatbot was appropriate" [28]

We also asked participants for free-text feedback describing *why* they liked or disliked the chatbot conversations ("*What did you like or dislike about your conversation with the chatbot?*").

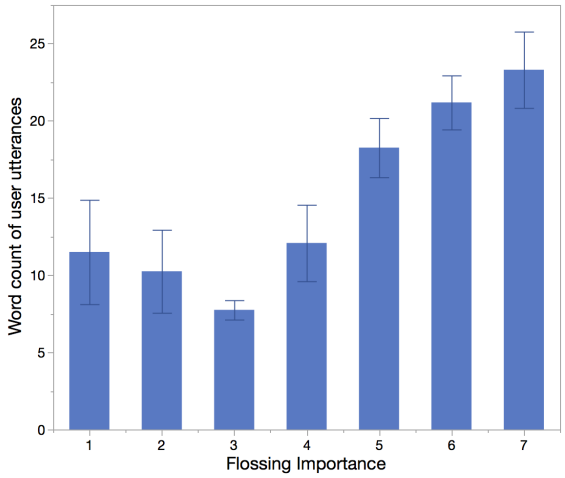


Figure 4: Users who perceived flossing as more important provided longer utterances.

5.5 Analysing User Utterances

We want to measure the quality of user utterances, but as utterances are brief (with each user providing two open-text responses), automated methods could not be used reliably. Consequently, we analysed the *response length* of user utterances, and manually labelled quality (in terms of *specificity* and *actionability* of utterances).

Response length was measured as the total word count of both user utterances. Word count has been used previously as an indicator for the quality of user utterances [53, 68, 86].

To label quality, we took inspiration from Xiao et al. and labelled utterances to give a human interpreted proxy for response quality [86, 87]. Xiao et al. were guided by Gricean maxims and labelled user utterances in terms of clarity, relevancy, and specificity. Due to the clarity and relevancy of user utterances being consistently good (with no responses being gibberish or irrelevant), we chose to label only for the maxim of specificity. From collected utterances, we want users to identify specific flossing barriers, and actionable strategies to overcome barriers. Consequently, we labelled barrier identifying utterances in terms of **specificity** and strategy identifying utterances in terms of **actioability**.

5.5.1 *Labelling Specificity of Utterances.*

When labelling specificity of barriers, we used the following guidelines:

> "*Does it give you enough info on how to identify and then give advice on how to overcome the barrier? For example, "I forget" is less specific and more difficult to act on, whereas "I forget when I'm..." is a more specific response.*"

This gave us three labels:

- "***0 (bad)** - no barrier is identified in the utterance*" such as the utterance: "*yes*".
- "***1 (Ok)** - a barrier is alluded to, but the utterance is not specific enough to allow for detailed guidance on how to overcome the barrier*" such as the utterance: "*I forget to do it regularly*".
- "***2 (good)** - a barrier is described in a detailed enough way to provide guidance on how to overcome the barrier*" such as the utterance: "*Simply remembering or finding the floss since I hide it from my child so it doesn't get destroyed.*".

5.5.2 *Labelling Actionability of Utterances.*

When labelling actionability of strategies, we used the following guidelines:

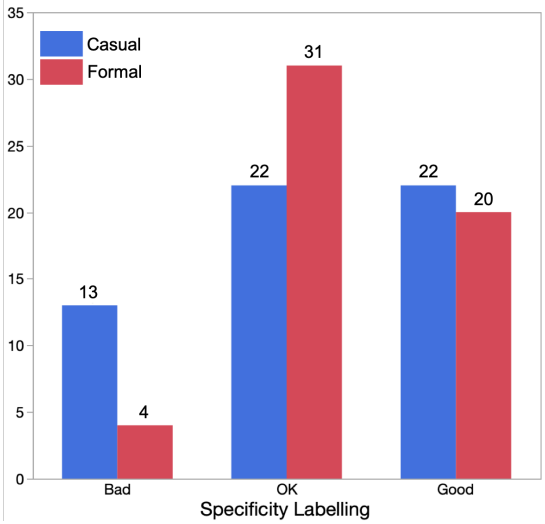
> "*Is the strategy described concrete enough to act upon? (Is it actionable?). A clear strategy can be followed such as "set an alarm" rather than something like "just do it"."*

This gave us three labels:

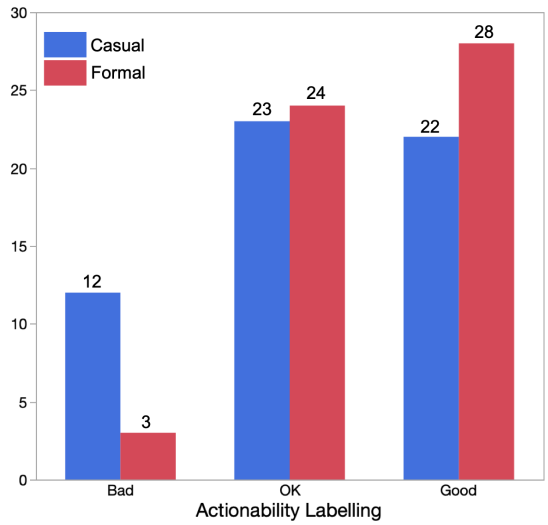
- "***0 (bad)** - No strategy is identified*" such as the utterance: "*I don't know*".
- "***1 (Ok)** - A strategy is alluded to, but may be abstract, under-explained or difficult to act upon*" such as the utterance: "*make it easier*".
- "***2 (good)** - A strategy is identified and it is clear what is intended to be done*" such as the utterance: "*Maybe an external reminder, like an alarm on my phone or something like that.*"

Table 3: Outcome measures for Study 2 (shown as "Mean (s.d.)"). Significant differences shown in bold.

		Casual	Formal
Subjective Measures (7-point Likert)	Continue	4.06 (1.64)	4.01 (1.52)
	Enjoyment	4.45 (1.58)	4.42 (1.62)
	Warmth	**5.31 (1.31)**	**4.71 (1.47)**
	Competency	4.61 (1.64)	4.57 (1.48)
	Appropriateness	5.70 (1.07)	5.67 (1.19)
Response Quality Measures	Response length (words)	19.69 (12.78)	19.82 (16.13)
	Specificity of barriers Actionability of strategies	See Figure 5	



(a) Formal participants provided fewer "bad" and more "ok" specificity responses than Casual participants.



(b) Formal participants provided fewer "bad" actionability responses than Casual participants.

Figure 5: Count of specificity and actionability labelling for user utterances. Participants provide more specific and actionable utterances when conversing with a formal chatbot compared to a casual chatbot.

5.6 Post Interaction Survey

On experiment completion, participants answered a post-interaction survey measuring belief in robotic intelligence and robotic feelings (using measures described in Section 3.6), and perceptions of dental flossing (measured on 7-point Likert scales - 1 = Strongly Disagree to 7 = Strongly Agree).

We used 3 measures of flossing perception [9] that are influenced by the extended Health Belief Model [60], namely: adherence: "*I try to floss my teeth every day*", flossing importance: "*It is important to floss my teeth every day*", and self-efficacy "*It is easy to floss my teeth every day*".

6 STUDY 2: RESULTS AND FINDINGS

We fit a linear model on each of our dependent variables with conversational style as the fixed effect, and we performed post-hoc Tukey's HSD to identify any specific differences. See Table 3 for a summary of the results by conversational style. We will now discuss specific findings and support these with participant quotes.

Conversational style was only significantly different for perceived warmth ($p = 0.0096$), with Casual being perceived as more warm than Formal. This contrasts with Study 1 where Formal participants perceived a chatbot as more appropriate and competent when being asked for sensitive health information. This difference in results may be due to medical history requested in Study 1 being perceived as more sensitive than dental flossing behaviour in Study 2, and thereby requiring a more formal conversational style. Additionally, Study 2 participants may appreciate the increased warmth of Casual when they are discussing their own personal behaviour (rather than reflecting on hypothetical relaying of information in Study 1).

Participants who believed in robotic intelligence were more likely to find the chatbot enjoyable ($p = 0.0053$), warm ($p < .0001$), competent ($p < .0001$) and appropriate ($p = 0.0004$), while those who

believed in robotic feelings were more likely find the chatbot enjoyable (p = 0.0031).

Participants who perceived flossing as important had more desire to continue chatbot use (p = 0.0113), found conversations more enjoyable (p = 0.0025), found the chatbot's tone more appropriate (p = 0.0153), and wrote longer responses (p = 0.0006; see Figure 4) than those with low perceived dental flossing importance. These findings highlight the value of educating users on the importance of healthy behaviours, and are supported by the extended Health Belief Model [60] which describes perceived importance as a determinant of healthy behaviour.

6.1 Response Informativeness

Two members of the research team independently coded user utterances while blind to source condition. The intraclass correlation (ICC) for Specificity labels was 0.902, and the ICC for Actionability labels was 0.821 (above the "excellent" threshold of 0.75 from Cicchetti et al. [16]). We excluded utterances from participants who floss each day, as they commonly described having no barriers to flossing due to their current adherence to the healthy behaviour. Typical utterances for fully adhering participants were: "*No everything is smooth and easy!*". This excluded utterances from 44 participants (23 casual and 21 formal) leaving a total of 112 participants (57 casual and 55 formal).

For the remaining user utterances that were initially labelled identically between coders, Formal users were more likely to provide more specific ($\chi^2(2, N = 98) = 10.153$, p = .0062) and actionable utterances ($\chi^2(2, N = 84) = 12.210$, p = .0022). After discussing and resolving specificity and actionability labels differing between the two coders (see Figure 5), it was still found that Formal users were more likely to provide specific ($\chi^2(2, N = 112) = 6.612$, p = .0367) and actionable utterances ($\chi^2(2, N = 112) = 6.490$, p = .0390).

Interestingly, the driving force of the difference seems to be that a formal conversational style elicits less "bad" utterances compared to a casual style. These results indicate a formal style might have made participants exert more effort in providing their responses.

7 SUMMARY OF KEY FINDINGS:

In summary, our key findings for research questions one to three are as follows:

- **RQ1:** Users were least likely to disclose sensitive health information, followed by sensitive non-health, and were most likely to disclose non-sensitive non-health information.
- **RQ2:** We did not find a difference in conversational style and a user's likelihood to disclose. However, (when being asked for their medical history) users perceive a chatbot with a formal conversational style as more appropriate and competent.
- **RQ3:** Users who do not fully adhere to a healthy behaviour wrote higher quality utterances when prompted with a formal conversational style.

8 DISCUSSION

Here we discuss the implications of both Studies 1 and 2. Interpretations for study-specific findings can be found in Sections 4 and 6. Our goal was to investigate the effect of a chatbot's language formality on self-disclosure. Our findings provide some empirical evidence that a chatbot using a more formal conversational style could be beneficial within the health domain by improving perceptions of competency and appropriateness (Study 1), and by eliciting higher quality user utterances (Study 2).

These findings could be explained due to differences from the expectations held by users. The Expectancy Violations Theory states that interactions either conform to or violate a user's expectation of how an agent should behave [10]. Qualitative feedback from Study 1 shows that participants expect a chatbot to use a more formal style when requesting their health information, and similarly subjective scores show that users perceive a casual style more negatively (less appropriate and competent) when sensitive health information was being requested. This negative violation (an undesired and unexpected act) led users to rate the casual conversational style more harshly (as opposed to rating the formal style more highly). This finding highlights the importance of using an appropriate conversational style in relation to the sensitivity of the information being requested by the chatbot (with medical history in Study 1 being the least likely piece of information for users to disclose, and thus leading to expectations of a more formal conversational style). Designers should carefully choose an appropriate conversational style if sensitive information is being requested so as not to violate any user expectations.

Furthermore, greater quality elicitation in Study 2 may be due to a mirroring effect. Previous work has found that people may mirror the response length of their (chatbot) interlocutor [36], and Social Exchange Theory [26] describes how people reciprocate the level of effort within a conversation. Relevant to this, Casual users in Study 2 produced more low quality user utterances, which may be due to a similar mimicry effect, whereby users are emulating the less specific language style of the casual prompts. In this sense, users may perceive a casual style as being lower effort (or formal as being higher effort) and mimic the style accordingly. Study 2 results could imply that a formal conversational style should also be used when users are discussing their health habits with a chatbot. However, as the domain of the Study 2 chatbot did not go beyond discussing healthy habits, implications can not be generalised confidently.

These findings could be seen as opposing previous studies that found that higher quality (multiple-choice) responses are collected by chatbots using a casual conversational style [11, 43]. However, the baseline conversational style of these studies were (non-conversational) standarised surveys, rather than a formal conversational style. Therefore increases in response quality could be attributed to using a conversational style rather than a terse script. Additionally, Zhou et al. compared two chatbots that used ensembles of features to create high-level personalities. They found that during high-stakes job interviews users are more willing to confide in a serious and assertive chatbot [88]. This matches our findings that under certain contexts, a more appropriate conversational style is expected (which could therefore lead to better response quality).

However, Study 1 findings also indicate that (when requesting less sensitive information) a formal conversational style is not expected by users, and a casual style could be adopted without negatively violating user expectations. This suggests chatbot conversational style could then be tailored more flexibly to the brand

or user expectation of the chatbot. For example, a chatbot for children could use a more casual conversational style to sound more straight-forward and friendly. On from this, there is also evidence that some users found a formal conversational style less empathetic (anecdotally from Study 1 user quotes, Section 4.1) and warm (see Table 3 for Study 2 subjective ratings). From Study 1, this could imply that (when non-sensitive information is being requested by the chatbot) a more casual conversational style could be used if designers desire a more empathetic and human-like chatbot. Warmth measures being statistically higher for Casual in Study 2, could imply that people may appreciate a more warm chatbot when they are discussing their health behaviours (similarly to disclosing more to a family member if they are more warm due to lessened feelings of judgement [23, 39]). Adopting a more casual (and thereby warm) conversational style could then lead users to co-operate and interact with the chatbot for longer [42].

Building on this, it is perhaps more clear which style to use for chatbots with a narrow domain of expertise or clear user expecations. However for multi-expertise tasks (potentially involving multiple domains and levels of information sensitivity) the choice of conversational style could be more difficult. If a different domain may necessitate a different conversational style, a single agent switching conversational style could prove jarring [13, 15], but multi-chatbots could be used [14, 41, 64], whereby the user is referred to a (supposedly) different chatbot that adopts an appropriate conversational style.

Additionally, it could be satisfying to users if chatbots adopt a similar conversational style to the user themselves. The use of linguistic mimicry (adopting the language style of another individual) leads to feelings of empathy [61]; and within information-seeking conversations between people, style mismatch can lead to less satisfying conversations [78, 79]. Related to this, users found a voice-based conversational agent to be more trustworthy if it matched their own conversational style [37]; multi-lingual users preferred a chatbot that mimicked their own code-mixing language (using multiple dialects) [4]; and recent work has investigated the development of style matching chatbots [37, 69, 78]. On from this, the effect of style matching on a user's quality of self-disclosure to a chatbot could be investigated further.

In addition our work highlights the importance of a user's prior belief in chatbot capabilities, with both Studies 1 and 2 finding that people perceived chatbot interactions more favourably if they believed in robotic feelings or intelligence. This matches previous findings that people base their judgement of systems on their own pre-existing beliefs [33, 44, 65, 85]. While these findings might suggest that people should be primed into believing a chatbot will be intelligent, Khadpe et al. noted the reverse to be true. They found users perceived a chatbot more positively if they had low expectations of a chatbot's competency before interacting with it [42].

9 LIMITATIONS

We will now outline our studies' limitations. Study 1 elicited users' likelihood to disclose, which may differ from actual disclosure [89]. Additionally, in Studies 1 and 2 we asked participants to provide subjective ratings such as warmth, competence, enjoyability, and desire to continue; measures that may be difficult for participants to quantify without more extended and in-depth interactions with the chatbot.

Experiments were not field studies, rather we used crowdsourcing where participants may be less intrinsically motivated or behave differently to volunteers or field study participants. However, crowdworkers have been found to provide accurate data [38, 62, 66], and be representative of the US population [62].

Our Study 2 responses come from desktops or laptops, although one of the most popular mediums of communication is mobile phones. External validity could be affected if people give shorter responses on mobile [3, 55], or have different expectations for chatbots on mobile (e.g., a more casual style [43]).

We also cannot claim generality over different modalities (such as voice), or different levels of language formality (as we did not ask questions in a wide variety of formality levels, but rather two contrasting examples within a certain domain). Similarly, it may be difficult to generalise results beyond the domains of Studies 1 (finance) and 2 (low-sensitivity health behaviours). Results could vary over a longitudinal study, such as participants reacting negatively to messaging they find more repetitive [45, 84].

10 CONCLUSION

This study investigates how the language formality of a chatbot's utterances affect the quality of self-disclosure from a user. We reported the results of two user studies conducted over AMT. Our findings suggest a formal conversational style may be perceived more positively when a chatbot is requesting sensitive health information, and may elicit more high quality user utterances when discussing a user's health behaviour.

ACKNOWLEDGMENTS

This research is supported by the Sea-NExT Joint Lab.

REFERENCES

[1] Martin Adam and Johannes Klumpe. 2019. Onboarding with a Chat–The Effects of Message Interactivity and Platform Self-Disclosure on User Disclosure Propensity. *Proceedings of the 27th European Conference on Information Systems (ECIS)* (May 2019). https://aisel.aisnet.org/ecis2019_rp/68

[2] Odette Aguirre-Zero, C Westerhold, R Goldsworthy, and Gerardo Maupome. 2016. Identification of barriers and beliefs influencing engagement by adult and teen Mexican-Americans in oral health behaviors. *Community Dental Health* 33, 1 (2016), 44.

[3] Christopher Antoun, Mick P Couper, and Frederick G Conrad. 2017. Effects of mobile versus PC web on survey response quality: A crossover experiment in a probability web panel. *Public Opinion Quarterly* 81, S1 (2017), 280–306.

[4] Anshul Bawa, Pranav Khadpe, Pratik Joshi, Kalika Bali, and Monojit Choudhury. 2020. Do Multilingual Users Prefer Chat-bots that Code-mix? Let's Nudge and Find Out! *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–23.

[5] Timothy Bickmore and Justine Cassell. 2000. "How about this weather?" Social Dialogue with Embodied Conversational Agents. In *Proc. AAAI Fall Symposium on Socially Intelligent Agents*.

[6] Timothy Bickmore and Justine Cassell. 2001. Relational Agents: A Model and Implementation of Building User Trust. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 396–403.

[7] Prakhar Biyani, Kostas Tsioutsiouliklis, and John Blackmer. 2016. "8 Amazing Secrets for Getting More Clicks": Detecting Clickbaits in News Streams Using Article Informality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30.

[8] Penelope Brown, Stephen C. Levinson, and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage*. Cambridge University Press. Google-Books-ID: OG7W8yA2XjcC.

[9] Maria E Buglar, Katherine M White, and Natalie G Robinson. 2010. The role of self-efficacy in dental patients' brushing and flossing: testing an extended Health Belief Model. *Patient Education and Counseling* 78, 2 (2010), 269–272.

[10] Judee K Burgoon, Joseph A Bonito, Paul Benjamin Lowry, Sean L Humpherys, Gregory D Moody, James E Gaskin, and Justin Scott Giboney. 2016. Application of expectancy violations theory to communication with and judgments about embodied agents during a decision-making task. *International Journal of Human-Computer Studies* 91 (2016), 24–36.

[11] Irene Celino and Gloria Re Calegari. 2020. Submitting surveys via a conversational interface: An evaluation of user acceptance and approach effectiveness. *International Journal of Human-Computer Studies* 139 (July 2020), 102410. https://doi.org/10.1016/j.ijhcs.2020.102410

[12] Jesse Chandler, Cheskie Rosenzweig, Aaron J Moss, Jonathan Robinson, and Leib Litman. 2019. Online panels in social science research: Expanding sampling methods beyond Mechanical Turk. *Behavior Research Methods* 51, 5 (2019), 2022–2038.

[13] Ana Paula Chaves, Eck Doerry, Jesse Egbert, and Marco Gerosa. 2019. It's How You Say It: Identifying Appropriate Register for Chatbot Language Design. In *Proceedings of the 7th International Conference on Human-Agent Interaction*. 102–109.

[14] Ana Paula Chaves and Marco Aurelio Gerosa. 2018. Single or multiple conversational agents? An interactional coherence comparison. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.

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

[16] Domenic V Cicchetti. 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment* 6, 4 (1994), 284.

[17] Patricia MA Corby, Aaron Biesbrock, Robert Bartizek, Andrea L Corby, Robin Monteverde, Rafael Ceschin, and Walter A Bretz. 2008. Treatment outcomes of dental flossing in twins: molecular analysis of the interproximal microflora. *Journal of Periodontology* 79, 8 (2008), 1426–1433.

[18] Samuel Rhys Cox, Yunlong Wang, Ashraf Abdul, Christian von der Weth, and Brian Y. Lim. 2021. Directed Diversity: Leveraging Language Embedding Distances for Collective Creativity in Crowd Ideation. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–35. https://doi.org/10.1145/3411764.3445782

[19] Amy JC Cuddy, Susan T Fiske, and Peter Glick. 2008. Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. *Advances in Experimental Social Psychology* 40 (2008), 61–149.

[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:1306.6078 [physics]* (June 2013). http://arxiv.org/abs/1306.6078 arXiv: 1306.6078.

[21] Tamara Dinev and Paul Hart. 2006. An Extended Privacy Calculus Model for E-Commerce Transactions. *Information Systems Research* 17, 1 (2006), 61–80.

[22] Doris Dippold, Jenny Lynden, Rob Shrubsall, and Rich Ingram. 2020. A turn to language: How interactional sociolinguistics informs the redesign of prompt: response chatbot turns. *Discourse, Context & Media* 37 (2020), 100432.

[23] Aryn M Dotterer and Elizabeth Day. 2019. Parental knowledge discrepancies: Examining the roles of warmth and self-disclosure. *Journal of Youth and Adolescence* 48, 3 (2019), 459–468.

[24] Daniëlle Duijst. 2017. Can we improve the user experience of chatbots with personalisation. *Master's thesis. University of Amsterdam* (2017).

[25] Ela Elsholz, Jon Chamberlain, and Udo Kruschwitz. 2019. Exploring language style in chatbots to increase perceived product value and user engagement. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*. 301–305.

[26] Richard M Emerson. 1976. Social exchange theory. *Annual Review of Sociology* 2, 1 (1976), 335–362.

[27] Susan Tufts Fiske and Shelley E Taylor. 1991. Social Cognition, McGraw-Hill. *New York, NY* (1991).

[28] Asma Ghandeharioun, Daniel McDuff, Mary Czerwinski, and Kael Rowan. 2019. EMMA: An Emotion-Aware Wellbeing Chatbot. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 1–7.

[29] David G Gillam and Huda Yusuf. 2019. Brief Motivational Interviewing in Dental Practice. *Dentistry Journal* 7, 2 (2019), 51.

[30] Richard C Graves, Judith A Disney, and John W Stamm. 1989. Comparative effectiveness of flossing and brushing in reducing interproximal bleeding. *Journal of Periodontology* 60, 5 (1989), 243–247.

[31] Herbert P Grice. 1975. Logic and Conversation. In *Speech Acts, Volume 3 of Syntax and Semantics*. Brill, 41–58. https://doi.org/10.1163/9789004368811_003

[32] Jingya Guo, Jiajing Guo, Changyuan Yang, Yanjing Wu, and Lingyun Sun. 2021. Shing: A Conversational Agent to Alert Customers of Suspected Online-payment Fraud with Empathetical Communication Skills. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–11.

[33] Jan Hartmann, Antonella De Angeli, and Alistair Sutcliffe. 2008. Framing the User Experience: Information Biases on Website Quality Judgement. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 855–864.

[34] Michael Hecker and TS Dillon. 2006. Ontological privacy support for the medical domain. *eHPass* (2006).

[35] Francis Heylighen and Jean-Marc Dewaele. 1999. Formality of Language: definition, measurement and behavioral determinants. *Interner Bericht, Center "Leo Apostel", Vrije Universiteit Brüssel* 4 (1999).

[36] 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.

[37] Rens Hoegen, Deepali Aneja, Daniel McDuff, and Mary Czerwinski. 2019. An end-to-end conversational style matching agent. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*. 111–118.

[38] John J Horton, David G Rand, and Richard J Zeckhauser. 2011. The online laboratory: Conducting experiments in a real labor market. *Experimental Economics* 14, 3 (2011), 399–425.

[39] Nina Howe, Jasmin Aquan-Assee, William M Bukowski, Pascale M Lehoux, and Christina M Rinaldi. 2001. Siblings as confidants: Emotional understanding, relationship warmth, and sibling self-disclosure. *Social Development* 10, 4 (2001), 439–454.

[40] Marie-Claire Jenkins, Richard Churchill, Stephen Cox, and Dan Smith. 2007. Analysis of user interaction with service oriented chatbot systems. In *International Conference on Human-Computer Interaction*. Springer, 76–83.

[41] Soyoung Jung, Kwan Min Lee, and Frank Biocca. 2015. Voice control system and multiplatform use: Specialist vs. generalist?. In *International Conference on Human Interface and the Management of Information*. Springer, 607–616.

[42] Pranav Khadpe, Ranjay Krishna, Li Fei-Fei, Jeffrey T Hancock, and Michael S Bernstein. 2020. Conceptual Metaphors Impact Perceptions of Human-AI Collaboration. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–26.

[43] Soomin Kim, Joonhwan Lee, and Gahgene Gweon. 2019. Comparing data from chatbot and web surveys: Effects of platform and conversational style on survey response quality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[44] Kristen J Klaaren, Sara D Hodges, and Timothy D Wilson. 1994. The role of affective expectations in subjective experience and decision-making. *Social Cognition* 12, 2 (1994), 77–101.

[45] Rafal Kocielnik and Gary Hsieh. 2017. Send Me a Different Message: Utilizing Cognitive Space to Create Engaging Message Triggers *(CSCW '17)*. Association for Computing Machinery, 2193–2207. https://doi.org/10.1145/2998181.2998324

[46] Charles E Lance, Marcus M Butts, and Lawrence C Michels. 2006. The Sources of Four Commonly Reported Cutoff Criteria: What Did They Really Say? *Organizational Research Methods* 9, 2 (2006), 202–220.

[47] Yi-Chieh Lee, Naomi Yamashita, and Yun Huang. 2020. Designing a Chatbot as a Mediator for Promoting Deep Self-Disclosure to a Real Mental Health Professional. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–27.

[48] Yuting Liao and Jiangen He. 2020. Racial mirroring effects on human-agent interaction in psychotherapeutic conversations. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 430–442.

[49] Zhaojiang Lin, Peng Xu, Genta Indra Winata, Farhad Bin Siddique, Zihan Liu, Jamin Shin, and Pascale Fung. 2020. CAiRE: An End-to-End Empathetic Chatbot. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 13622–13623. https://doi.org/10.1609/aaai.v34i09.7098

[50] Courtney Linder. 2020. The Effects of a Healthcare Chatbots' Language and Persona on User Trust, Satisfaction, and Chatbot Effectiveness. *Master's thesis. Clemson University* (2020).

[51] Bingjie Liu and S Shyam Sundar. 2018. Should machines express sympathy and empathy? Experiments with a health advice chatbot. *Cyberpsychology, Behavior, and Social Networking* 21, 10 (2018), 625–636.

[52] Irene Lopatovska, Elena Korshakova, Diedre Brown, Yiqiao Li, Jie Min, Amber Pasiak, and Kaige Zheng. 2021. User perceptions of an intelligent personal assistant's personality: the role of interaction context. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*. 15–25.

[53] Stephen Louw, R Watson Todd, and P Jimarkon. 2011. Active listening in qualitative research interviews. In *Proceedings of the International Conference: Research in Applied Linguistics, April*.

[54] Ereni Markos, George R Milne, and James W Peltier. 2017. Information sensitivity and willingness to provide continua: a comparative privacy study of the United States and Brazil. *Journal of Public Policy & Marketing* 36, 1 (2017), 79–96.

[55] Aigul Mavletova. 2013. Data quality in PC and mobile web surveys. *Social Science Computer Review* 31, 6 (2013), 725–743.

[56] Youngme Moon. 2000. Intimate Exchanges: Using Computers to Elicit Self-Disclosure From Consumers. *Journal of Consumer Research* 26, 4 (2000), 323–339. http://www.jstor.org/stable/10.1086/209566

[57] Clifford Nass, Jonathan Steuer, and Ellen R Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 72–78.

[58] Scott Nowson, Jon Oberlander, and Alastair J. Gill. 2005. Weblogs, genres and individual differences. In *In Proceedings of the 27th Annual Conference of the Cognitive Science Society*. 1666–1671.

[59] Daniel M Oppenheimer, Tom Meyvis, and Nicolas Davidenko. 2009. Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology* 45, 4 (2009), 867–872.

[60] Rita Orji, Julita Vassileva, and Regan Mandryk. 2012. Towards an Effective Health Interventions Design: An Extension of the Health Belief Model. *Online Journal of Public Health Informatics* 4, 3 (2012).

[61] Jahna Otterbacher, Chee Siang Ang, Marina Litvak, and David Atkins. 2017. Show me you care: Trait empathy, linguistic style, and mimicry on Facebook. *ACM Transactions on Internet Technology (TOIT)* 17, 1 (2017), 1–22.

[62] Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. 2010. Running Experiments on Amazon Mechanical Turk. *Judgment and Decision Making* 5, 5 (2010), 411–419.

[63] W Barnett Pearce and Stewart M Sharp. 1973. Self-Disclosing Communication. *Journal of Communication* 23, 4 (1973), 409–425.

[64] Claudio S Pinhanez, Heloisa Candello, Mauro C Pichiliani, Marisa Vasconcelos, Melina Guerra, Maíra G de Bayser, and Paulo Cavalin. 2018. Different but equal: Comparing user collaboration with digital personal assistants vs. teams of expert agents. *arXiv preprint arXiv:1808.08157* (2018).

[65] Eeva Raita and Antti Oulasvirta. 2011. Too good to be bad: Favorable product expectations boost subjective usability ratings. *Interacting with Computers* 23, 4 (2011), 363–371.

[66] David G Rand. 2012. The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology* 299 (2012), 172–179.

[67] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y.-Lan Boureau. 2019. Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset. *arXiv:1811.00207 [cs]* (Aug. 2019). http://arxiv.org/abs/1811.00207 arXiv: 1811.00207.

[68] Jungwook Rhim, Minji Kwak, Yeaeun Gong, and Gahgene Gweon. 2022. Application of humanization to survey chatbots: Change in chatbot perception, interaction experience, and survey data quality. *Computers in Human Behavior* 126 (2022), 107034.

[69] Hannes Ritschel, Andreas Seiderer, Kathrin Janowski, Stefan Wagner, and Elisabeth André. 2019. Adaptive linguistic style for an assistive robotic health companion based on explicit human feedback. In *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments (PETRA)*. 247–255.

[70] Kambiz Saffarizadeh, Maheshwar Boodraj, and Tawfiq M Alashoor. 2017. Conversational Assistants: Investigating Privacy Concerns, Trust, and Self-Disclosure.. In *Thirty Eighth International Conference on Information Systems (ICIS)*.

[71] Benjamin Schüz, Falko F Sniehotta, Amelie Wiedemann, and Rainer Seemann. 2006. Adherence to a daily flossing regimen in university students: Effects of planning when, where, how and what to do in the face of barriers. *Journal of Clinical Periodontology* 33, 9 (2006), 612–619.

[72] Vasant Srinivasan and Leila Takayama. 2016. Help me please: Robot politeness strategies for soliciting help from humans. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 4945–4955.

[73] S Shyam Sundar and Jinyoung Kim. 2019. Machine heuristic: When we trust computers more than humans with our personal information. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–9.

[74] S Shyam Sundar, Jinyoung Kim, Mary Beth Rosson, and Maria D Molina. 2020. Online privacy heuristics that predict information disclosure. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[75] Ruby Suresh, Katharine C Jones, Jonathan Timothy Newton, and Koula Asimakopoulou. 2012. An exploratory study into whether self-monitoring improves adherence to daily flossing among dental patients. *Journal of Public Health Dentistry* 72, 1 (2012), 1–7.

[76] Laura Teddiman. 2009. Contextuality and beyond: Investigating an online diary corpus. In *Third International AAAI Conference on Weblogs and Social Media*.

[77] Indrani Medhi Thies, Nandita Menon, Sneha Magapu, Manisha Subramony, and Jacki O'neill. 2017. How do you want your chatbot? An exploratory Wizard-of-Oz study with young, urban Indians. In *IFIP Conference on Human-Computer Interaction*. Springer, 441–459.

[78] Paul Thomas, Mary Czerwinksi, Daniel McDuff, and Nick Craswell. 2021. Theories of conversation for conversational IR. *ACM Transactions on Information Systems (TOIS)* 39, 4 (2021), 1–23.

[79] Paul Thomas, Mary Czerwinski, Daniel McDuff, Nick Craswell, and Gloria Mark. 2018. Style and alignment in information-seeking conversation. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*. 42–51.

[80] Amos Tversky and Daniel Kahneman. 1974. Judgment Under Uncertainty: Heuristics and Biases. *Science* 185, 4157 (1974), 1124–1131.

[81] Alastair van Heerden, Xolani Ntinga, and Khanya Vilakazi. 2017. The potential of conversational agents to provide a rapid HIV counseling and testing services. In *2017 International Conference on the Frontiers and Advances in Data Science (FADS)*. IEEE, 80–85.

[82] Marilyn A Walker, Janet E Cahn, and Stephen J Whittaker. 1997. Improvising linguistic style: Social and affective bases for agent personality. In *Proceedings of the first international conference on Autonomous agents*. 96–105.

[83] Thiemo Wambsganss, Rainer Winkler, Matthias Söllner, and Jan Marco Leimeister. 2020. A Conversational Agent to Improve Response Quality in Course Evaluations. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–9.

[84] David Westerman, Autumn P Edwards, Chad Edwards, Zhenyang Luo, and Patric R Spence. 2020. I-It, I-Thou, I-Robot: The Perceived Humanness of AI in Human-Machine Communication. *Communication Studies* (2020), 1–16.

[85] Daricia Wilkinson, Öznur Alkan, Q Vera Liao, Massimiliano Mattetti, Inge Vejsbjerg, Bart P Knijnenburg, and Elizabeth Daly. 2021. Why or why not? The effect of justification styles on chatbot recommendations. *ACM Transactions on Information Systems (TOIS)* 39, 4 (2021), 1–21.

[86] Ziang Xiao, Michelle X Zhou, Wenxi Chen, Huahai Yang, and Changyan Chi. 2020. If I Hear You Correctly: Building and Evaluating Interview Chatbots with Active Listening Skills. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[87] Ziang Xiao, Michelle X Zhou, Q Vera Liao, Gloria Mark, Changyan Chi, Wenxi Chen, and Huahai Yang. 2020. Tell me about yourself: Using an AI-powered chatbot to conduct conversational surveys with open-ended questions. *ACM Transactions on Computer-Human Interaction (TOCHI)* 27, 3 (2020), 1–37.

[88] Michelle X Zhou, Gloria Mark, Jingyi Li, and Huahai Yang. 2019. Trusting virtual agents: The effect of personality. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 9, 2-3 (2019), 1–36.

[89] Jakub Złotowski, Kumar Yogeeswaran, and Christoph Bartneck. 2017. Can we control it? Autonomous robots threaten human identity, uniqueness, safety, and resources. *International Journal of Human-Computer Studies* 100 (2017), 48–54.