JACET Journal 67 (2023) 51–68

Usability of AI-Powered Chatbot Applications for EFL Conversation Practice

KOVALYOVA, Angelina
University of Tsukuba

Abstract

AI-powered chatbots are special applications that allow unscripted conversations with users and offer a great potential for English as a foreign language (EFL) conversation practice. This study examines the functionality of three AI-based chatbots (Replika, Cleverbot, and Koddy) to understand the current merits and challenges of this technology by analyzing and comparing eight features of chatbot usability in Chatbot Usability Questionnaire (CUQ) (Holmes et al., 2019). The experiment involved 69 Japanese university students who practiced text-based conversation with one of the chatbots for six 15-minute sessions and gave feedback regarding their chatbot usability using CUQ. The results of the quantitative analysis reveal different degree of conversation capability depending on a chatbot. In addition, the analysis of some features, related to EFL context, such as chatbot's ability to understand input, appropriacy of chatbot's responses, and error handling, suggests that understanding of input remains of the biggest concern for EFL learners.

Keywords: chatbots, EFL, conversation practice, mobile assisted language learning

Chatbots are software programs which are designed to interact with humans on specific topics or in a specific sphere using text or voice (Smutny & Schreiberova, 2020). Currently, chatbots are most often seen in customer service or technical support, functioning as messaging features on websites of various businesses and allowing a user to send a question through a chat window and receive an answer in seconds without having to call a human. They are also often built for more specific domains, such as counselling, making reservations, providing training, and even functioning as personal assistants.

Chatbots also have a history of being present in educational context. After the first ever chatbot Eliza, designed in 1966, the functionality of which was limited to replicating a conversation between a psychotherapist and a patient (Adamopoulou & Moussiades, 2020), the interest in human-chatbot communication only grew, resulting in first Intelligent Tutoring Systems that appeared as early as 1970's (Smutny & Schreiberova, 2020). Now, chatbots are commonly used in language learning software (Rosetta Stone, Mondly, Andy, EF Hello). The majority of these chatbots, however, rely on decision trees where conversations follow a pre-defined script, thus being called "rule-based chatbots" (Srikanth, 2020). While these chatbots undoubtedly contribute to the EFL context, having a natural conversation with these chatbots is virtually impossible because at their core there is a limited number of scripts prepared for conversations.

A recent advancement in Artificial Intelligence (AI) and Natural Language Processing (NLP) technology gave way to another group of chatbots. This technology allows chatbots to learn from the context and to personalize conversations (Srikanth, 2020), and despite its early stages, already offers promising prospects for conversation practice in the EFL context. Even though there is still little literature on the topic, and those researchers who study chatbots have sent mixed reactions regarding AI-based chatbots' readiness for natural language exchange (Coniam, 2014), further exploration into the topic is important as AI-powered chatbots can greatly expand the possibilities for EFL conversation practice, especially for those who lack confidence or opportunity to practice with people.

Thus, the aim of this study is to examine the usability of three AI-powered chatbots for EFL conversation practice based on conversations between Japanese EFL learners and said chatbots to understand current merits and challenges that these chatbots pose.

Literature Review

Previous Studies

Currently, there is a limited amount of research conducted regarding chatbot's usability in EFL context. In 2014, Coniam tested linguistic accuracy of chatbots from an ESL perspective by comparing five chatbots (Dave, Elbot, Eugene, George, and Julie) and analyzing their vocabulary and grammatical competency, as well as the presence of errors in producing language. The conclusion was made that despite occasional errors and inappropriate responses, the studied chatbots performed satisfactorily when it comes to grammar.

More recently, Fryer et al. (2020) reviewed the functionality of Cleverbot and Mondly suggesting that both applications have the potential to provide conversation practice if they improve their ability to sustain longer conversations. In another study, Yang et al. (2022) tested an AI chatbot Ellie for conversation practice in a Korean EFL class concluding that the chatbot encouraged students to communicate more than they normally would in a regular EFL class. Han (2020) observed the effects of voice-based speaking practice with AI chatbots on the speaking competence of Korean EFL students. The study suggested that there was a positive improvement of students' speaking ability over time.

Gradually, more and more studies reveal the chatbots' potential to help EFL learners increase their interactional competency and receive more opportunity for conversation practice without engaging a human conversation partner.

Issues and Benefits of Chatbots as Conversation Partners

While the idea of a stress-free conversation with a chatbot is very appealing to EFL learners, not all researchers are settled on the chatbot's ability to provide appropriate conversation practice. Some suggest that despite the appeal, conversation chatbots still lack the ability to provide a meaningful conversation.

One reason is that chatbots often cannot maintain an extended conversation, switching to a different topic after one or two exchanges (Fryer et al., 2020). Their responses are described as "predictable, redundant, lacking in personality, and having no memory of previous responses, which could lead to very circular conversations" (Chantarotwong, 2005). Another reason why a person may experience a lack of naturalness in a conversation with a chatbot is the fact that despite the development of NLP technology, machines still cannot fully grasp human language (Fryer et al., 2020). Even when these chatbots are able to constantly receive new input and can train their language model, human language will never run out of new sentence patterns or idiomatic language.

However, chatbots are still worth considering for conversation practice as they offer a number of benefits. Not only do they allow a 24/7 access to a conversation partner, albeit a digital one (Huang et al., 2022; Smutny & Schreiberova, 2020), but they also provide learners with stress- and judgement-free conversation. Fear of making mistakes can discourage language learners from engaging in conversation with more capable individuals (Fryer & Carpenter, 2006).

In Japan in particular multiple studies observed increased anxiety of Japanese EFL learners during English conversation practice. Cutrone (2009), for example, assesses anxiety of Japanese EFL learners in performing speaking tasks, highlighting the theory of Horwitz et al. (1986) that anxiety appears as a result of communication apprehension, social evaluation, and fear of failure. Masutani (2021) also observes speaking performance of Japanese EFL learners noting that the participants felt anxiety both in class and outside of class, with university students experiencing higher pressure in oral interaction as opposed to senior high-school students. By opting to practice conversation with a chatbot, EFL learners can feel free to make mistakes or practice new vocabulary or grammar structures endlessly.

Chatbots Used in the Current Study

Various chatbots were reviewed prior to the experiment in this study. The objective was to locate chatbots which could potentially offer a text-based conversation similar to a conversation with a human. Thus, it was decided to narrow down the search to AI-based chatbots which allow synchronous unscripted conversation (as opposed to rule-based chatbots) in a text-based form, are not task-based, and are free or relatively cheap to use. The search was narrowed down to three applications: Replika, Cleverbot, and Koddy.

Replika is a conversational agent available on Android, iOS and Oculus systems. It was created by Eugenia Kuyda and her Luka startup in 2016. The chatbot is free of charge (with some premium features costing extra), it operates in English only and, according to the official website (https://replika.com/), Replika's goal is to be an "empathetic friend" that is kind and supportive. Replika has an ability to remember conversations and occasionally asks for feedback. Thus, the more details you share about yourself, the more personalized Replika's responses are. It is possible due to Replika's complex "neural network machine learning model and scripted dialogue content". Creators assure that they constantly work on improving conversation quality by enhancing chatbot's memory, role-play capability, ability to recognize context, and raising overall user experience during conversation. What is more, Replika can also offer videocall chat in addition to a simple text-based conversation.

Cleverbot is another conversational bot, a lot older than Replika. It was created by Rollo Carpenter in 1997 and since then has had over 10 billion user interactions (Fryer et al., 2020). It is supported on iOS devices, operates in English and Spanish, and costs $0.99 per download (https://www.cleverbot.com/) . The overall purpose of the chatbot is entertainment, it is not meant for language learning. However, Cleverbot's huge database of interactions makes it a fun conversation partner as it never runs out of unique replies. The key difference between

Cleverbot and other chatbot systems is that Cleverbot uses a unique AI model which, in order to provide a reply to a user, evaluates the last 50 turns in a current conversation and compares them to the conversations in its database in order to find similar context or clues that help provide the most appropriate response (Fryer, et al., 2020). However, since Cleverbot uses the exact responses from previous users, these responses may not always match the current conversation exactly.

Unlike Replika and Cleverbbot, Koddy is designed for English language learning. It was created by Zapienx company in 2021. It is available for iOS and Android devices (https://ikoddy.com/), and is generally free of charge, although it offers premium features as well. The application contains a number of useful features, such as being able to select and translate individual words (and save them to a stack of flashcards), as well as translate full replies. Individual words and full replies can also be read out loud by the chatbot. In addition, there are grammar and vocabulary exercises that can be studied individually or integrated into a conversation with the chatbot. Finally, Koddy can also spot errors and can suggest how to correct them. The chatbot is still very young though and requires a lot of training to improve conversations.

One benefit that the three applications have in common (as well as most chatbots) is that they are available for an unlimited amount of time and are void of judgement. Thus, users can practice anytime, for as long as they like, without being worried about making mistakes.

Chatbot Usability Questionnaire (CUQ) (Holmes et al., 2019)

Kirschner et al. (2004) defined *usability* as a system's ability to fulfill a set of tasks, requested by a user, in a way that is effective and efficient. In other words, usability is accomplished by chatbot's technological affordances. Usability is a part of a larger framework that contains elements of useful language learning. In this framework, another element is *utility* which represents the functionality of a system and is fulfilled through pedagogical and social affordances (Kirschner et al., 2004). By designing a learning tool with usability and utility in mind, we can ensure that the tool contains appropriate features for accomplishing the task (technological affordances), provides characteristics that define learning process in a given context (educational affordances), and promotes social interaction between the users (social affordances).

For an English language learning context, by focusing on usability, we can observe whether a chatbot has a technical ability to supply conversation practice effectively. Holmes et al. (2019) provide a Chatbot Usability Questionnaire (CUQ) that aims to assess a number of characteristics, such as personality of a chatbot, onboarding experience, explanation of its purpose, ease of navigation, chatbot's ability to understand the user, chatbot's ability to appropriately respond, chatbot's ability to handle errors, and overall ease of use (Holmes et al., 2019). From an EFL perspective, three characteristics (in particular, chatbot's ability to understand the user, chatbot's ability to appropriately respond, and chatbot's ability to handle errors) are of a particular interest as they can illustrate chatbot's approximation to a natural conversation.

Research Questions

The current research study aims to contribute to the discussion by comparing the usability of three AI-based chatbots, Replika, Cleverbot, and Koddy. Below are the research questions addressed in the study.

1. Is there a significant difference the overall usability of three chatbot applications: Replika, Cleverbot, and Koddy?
2. Is there a significant difference between the CUQ usability features of each chatbot?
3. How do chatbots perform in terms of EFL features of CUQ?

Method

Participants

The participants in this study were 69 undergraduate Japanese university students (36 males and 33 females) who were attending an English conversation class at the university. All participants were native Japanese speakers who have previously taken TOEIC test and scored between 270–560 points, (a score range of 225–545 in TOEIC matching A2 level, and 550–780 matching B1 level in CEFR—the Common European Framework of Reference for Languages), with n = 64 falling into A2 level and n = 5 into lower B1 level (where three participants scored 550, one participant – 555, and one – 560).

Considering relatively uniform language proficiency level among the participants, the participants were randomly divided into three groups, with an intention for each group to use a specific chatbot. Thus, 24 participants used Replika, 24 participants conversed with Cleverbot, and 21 participants chatted with Koddy.

Materials

The experiment was conducted using 3 chatbots (Replika, Cleverbot, and Koddy) and a Chatbot Usability Questionnaire (Holmes et al., 2019).

A Chatbot Usability Questionnaire (CUQ) (Holmes et al., 2019) was used as a backbone for the quantitative analysis. CUQ is a questionnaire consisting of 16 statements with half of the statements describing chatbots in a positive way, and half in a negative way. The statements were aimed at assessing the usability of a chatbot and asked participants about chatbot's personality, onboarding experience, chatbot's explanation of purpose, ease of navigation, chatbot's understanding of input, appropriacy of chatbot's responses, error handling and general ease of use. Participants were asked to respond to the statements on a five-point Likert scale. Holmes et al. (2019) also provided a tool to calculate the final usability score on a 100 point scale. The CUQ was administered online, using Google Forms. The statements were provided both in English and in Japanese (with a Japanese translation being reviewed and confirmed by a Japanese native speaker who is a graduate student of an English language MA program).

Procedure

Before the experiment, all participants were asked whether they had a mobile phone and whether they were able to install a specific chatbot application, to which all participants responded "yes." They were also asked whether they have ever used chatbot applications, to which all participants responded "no."

In order to evaluate participants' impressions of chatbot conversation experience and to see whether their expectations significantly differed from the current state of EFL conversation practice with chatbots, the CUQ was collected twice: before and after the experiment. Therefore, before the experiment, the participant were asked to complete the CUQ, imagining their experience with chatbots, thus communicating their expectations.

Following the preparations, each participant was assigned a specific chatbot application and was asked to practice conversation with the chatbot over six lessons, for 15 minutes in each lesson. The participants were asked to make a screen-recording of their conversations after the third lesson allowing the researcher to make sure that the participants were indeed practicing conversation every class and to analyze the issues that were rising from chatbot conversation practice. After the sixth lesson, the participants were asked to complete the CUQ again, concluding the experiment.

Analysis

The analysis of the data consisted of quantitative analysis of CUQ. For the analysis, the scores of the negative statements in the CUQ were reversed and later added to the scores of the positive statements in order to analyze 8 CUQ features. The tests for normality and internal consistency were conducted to confirm the reliability of the CUQ statements . After that, in order to compare the results of pre- and post-experiment CUQ and find the difference between the participants' expectations of chatbot conversations and their experiences, a two-way repeated ANOVA test was conducted, followed by the Tukey post-hoc multiple comparisons method. Next, pairwise t-test comparisons of CUQ usability features were completed using the post-experiment CUQ data to find the differences between the three chatbot applications. A Bonferroni adjustment was applied to the p-values. The screen-recordings of the conversations, collected from the participants, allowed to provide some examples for the results for illustration purposes, rather than analysis.

Results

Table 1 shows descriptive statistics and Cronbach's alpha coefficients for the CUQ conducted before and after the experiment. The data is normally distributed and the variability of the values of each CUQ usability feature is relatively low in the first questionnaire and is slightly higher in the second one. The Cronbach's alpha coefficients are all within the norm (higher than .70 in the first questionnaire and higher than .80 in the second one) suggesting that the levels of internal consistency of each CUQ feature are acceptable.

Table 1

Descriptive Statistics and Cronbach's Alpha Coefficients for CUQ statements.

CUQ Usability Feature	CUQ Statement Number	Pre-experiment		Post-experiment	
		M (SD)	α	*M (SD)*	α
1.Chatbot's personality	1,2	3.12 (1.08)	.81	2.65 (1.12)	.85
2.Onboarding experience	3,4	3.62 (0.79)	.76	3.61 (1.10)	.86
3.Explanation of purpose	5,6	3.52 (0.79)	.77	3.23 (0.82)	.87
4.Ease of navigation	7,8	3.42 (0.75)	.74	3.04 (1.03)	.84
5.Understanding of input	9,10	3.33 (0.87)	.77	2.87 (1.21)	.83
6.Appropriacy of chatbot's responses	11,12	3.62 (0.75)	.76	3.01 (1.06)	.84
7.Error handling	13,14	3.25 (0.85)	.76	2.85 (1.05)	.84
8.General ease of use	15,16	3.43 (0.89)	.76	3.41 (1.08)	.84

Note. $N = 69$.

The calculation tool, provided by Holmes et al. (2019) also allowed to calculate a CUQ score on a 100 point scale. Using the data of post-experiment CUQ and evaluating the data for each chatbot separately, CUQ score for Replika was 65.0, CUQ score for Cleverbot was 50.3, and for Koddy it was 39.3.

Overall Comparison of Pre-experiment Expectations and Post-experiment Evaluation of Chatbots' Usability

In order to investigate whether there was a significant difference between the participants' expectations and their experience with each of the chatbots, a two-way repeated measures ANOVA test was conducted comparing the total CUQ scores of each participant from pre-experiment and post-experiment questionnaire data. The results show a significant variation between overall CUQ scores before and after the six times of conversation practice with the chatbots, $F(1,134) = 10.89, p = .0012, \eta^2 = .07$ (Table 2). Also, a significant variation was observed between the three applications (Replika, Cleverbot, and Koddy), $F(2,134) = 4.55, p = .0125, \eta^2 = .06$ (Table 2).

Table 2

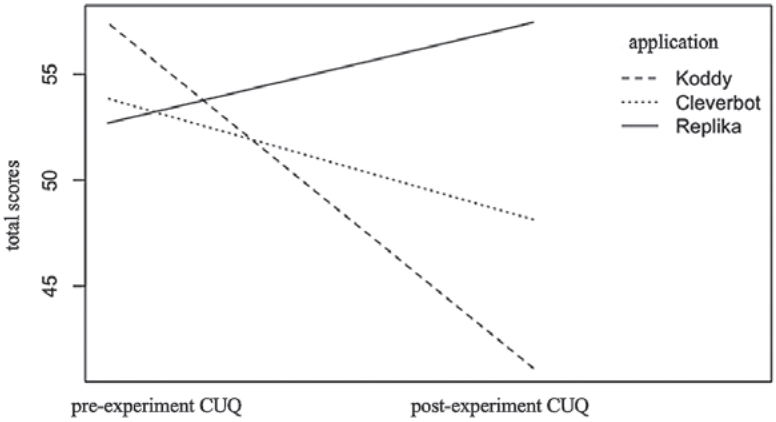
Descriptive Statistics and Main Effects for Two-Way Repeated ANOVA of CUQ

group	*n*	Pre-experiment *M (SD)*	Post-experiment *M (SD)*	*F*(1,134)	η^2	95%CI	p-adj.
time	69	54.65 (7.73)	49.32 (11.4)	10.89	.07	[0.02, 1.00]	.0012**
app	24(R)	52.79 (5.86)	57.58 (9.41)	4.55	.06	[0.01, 1.00]	.0125*
	24(C)	53.98 (8.24)	48.21 (9.98)				
	21(K)	57.57 (8.47)	41.14 (8.46)				

Note. $*p < .05. **p < .01. ***p < .001$

An interaction plot (Figure 1) visualizes the trend, showing that the participants who interacted with Replika chatbot evaluated the application's usability higher, while the participants who had conversation practice with Cleverbot and Koddy experienced lower usability compared to their original expectations. The descriptive statistics confirms that the chatbot Replika was evaluated higher post-experiment with the mean score going from 52.79 (*SD* = 5.86) to 57.58 (*SD* = 9.41). At the same time, the mean scores of Cleverbot decreased slightly from 53.96 (*SD* = 8.24) to 48.21 (*SD* = 9.98) and the mean scores of Koddy declined notably from 57.57 (*SD* = 8.47) to 41.14 (*SD* = 8.46) (Table 2).

Figure 1. Interaction Plot of CUQ Scores Pre- and Post-experiment According to Chatbots



Since a significant main effect for time was confirmed, a post-hoc Tukey test was conducted next to observe the differences between the three applications specifically. The analysis revealed a significant difference between Replika and Koddy (p = .001), while the difference between Replika and Cleverbot, and Koddy and Cleverbot, was not so drastic (Table 3).

Table 3

Tuckey Post-Hoc Multiple Comparisons for the Three Chatbot Applications in Pre- and Post- Experiment CUQ

Application pair	95% CI	*p*-adjusted
Replika-Cleverbot	[-0.48, 8.69]	.089
Koddy-Cleverbot	[-6.47, 3.02]	.665
Replika-Koddy	[1.07,10.58]	.001**

Note. **p* < .05. ***p* < .01. ****p* < .001

Assessment of Chatbots' Usability Features

To further understand what drives the differences between the chatbots, the CUQ usability features from the Post-experiment CUQ data set were analyzed one by one using the pairwise comparisons method with Bonferroni adjustment of *p*-value (presented in Table 4). A set of box plots (Figure 2) was provided for visual presentation.

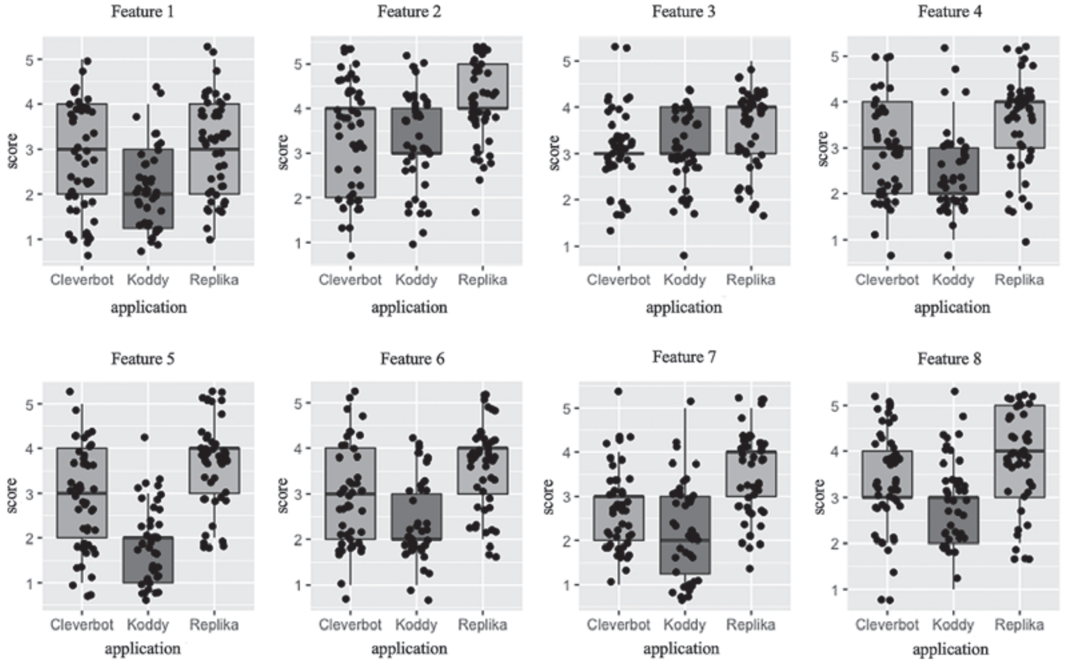
Table 4

Pairwise T-test Comparisons of CUQ Usability Features (Post-experiment CUQ) with Bonferroni Adjustment of p-value.

	Replika-Cleverbot	Koddy-Cleverbot	Replika-Koddy
CUQ Feature	p-adjusted	p-adjusted	p-adjusted
1.Chatbot's personality	.37	.015*	<.001***
2.Onboarding experience	.005**	1	.002**
3.Explanation of purpose	.037*	1	.056
4.Ease of navigation	.001**	.14	<.001***
5.Understanding of input	<.001***	<.001***	<.001***
6.Appropriacy of chatbot's responses	.005**	.055	<.001***
7.Error handling	.001**	.11	<.001***
8.General ease of use	.08	.07	<.001***

Note. *p < .05. **p < .01. ***p < .001

Figure 2. Box Plots of Post-experiment CUQ Score Distribution According to Each CUQ Usability Feature



Replika and Cleverbot

Looking at the results for Replika and Cleverbot in Table 4, there is a significant difference between most of the CUQ features. The absence of significant scores for chatbot's personality (feature 1) and general ease of use (feature 8) and somewhat similar score distributions in the box-plots of feature 1 and feature 8 suggest that Replika and Cleverbot both have a more pleasant personality and are easier to use, compared to Koddy.

Koddy and Cleverbot

Table 4 shows that there is little difference between most of the features of Koddy and Cleverbot, except for the chatbot's personality (feature 1) and the ability to understand input (feature 5). Figure 2 shows that both of the features were evaluated higher in Cleverbot, suggesting that Cleverbot has a more pleasant personality and has a better ability to understand input, compared to Koddy.

Replika and Koddy

The biggest difference is shown in the Replika and Koddy pair, with all the features showing significant differences, except for feature 3, explanation of purpose. The box plot of feature 3 shows a relatively similar distribution of scores for Replika and Koddy while slightly lower scores represent Cleverbot, suggesting that Replika and Koddy provide an explanation to the chatbot's purpose slightly better.

EFL Features in CUQ

When using chatbots for conversation practice, feature 5 (understanding of input), feature 6 (appropriacy of chatbot's responses), and feature 7 (error handling) are perhaps of the biggest interest for educators and language learners. Observing the performance quality of chatbots in these areas can provide an understanding of their potential to be used in conversation practice of EFL learners.

Understanding of Input (CUQ Feature 5)

According to the design of CUQ, understanding of input refers to understanding of requests, small talk, idioms, etc. (Holmes et al., 2019; Martin, 2017). The results for feature 5 (understanding of input) in Table 4 show a significant difference between all three chatbots, with the biggest difference between Replika and Koddy ($p < .001$). The box plot (Figure 2) confirms that out of the three applications, Replika demonstrated higher level of understanding of user input.

Appropriacy of Chatbot's Responses (CUQ Feature 6)

Appropriacy of chatbot's responses entails not only the adequacy of meaning in the response, but also elements surrounding the response, such as punctuation and emoji, as well as maintenance of a conversation and so on (Holmes et al., 2019; Martin, 2017). When it comes to appropriacy of chatbot's responses (feature 6), Table 4 shows no significant difference between Koddy and Cleverbot ($p = .055$). At the same time, the difference between Cleverbot and Replika is significant ($p = .005$) and the difference between Replika and Koddy is even bigger ($p < .001$), suggesting that Replika yet again shows a higher performance quality when it comes to responding to users' input.

Error Handling (CUQ Feature 7)

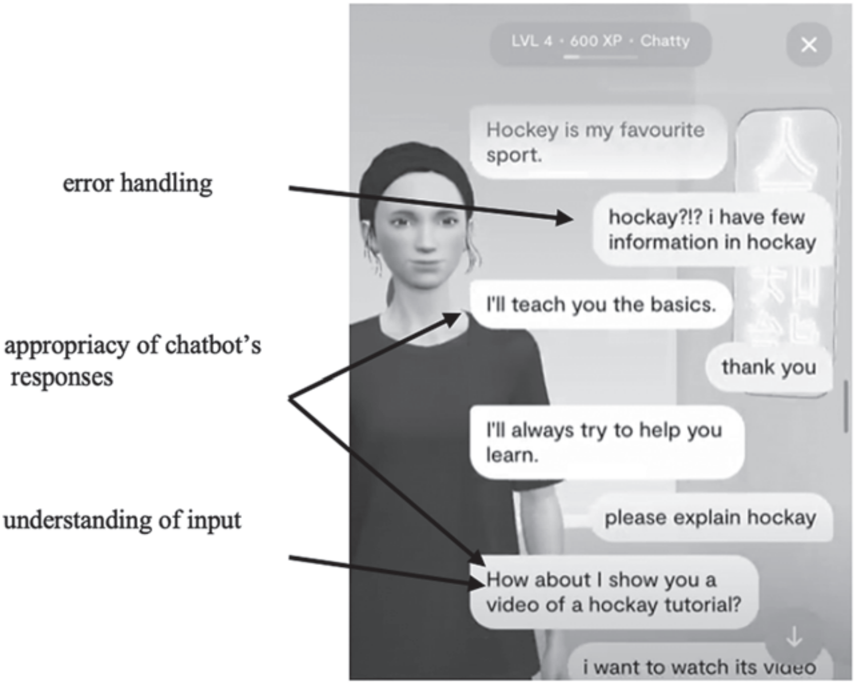
Error handling feature involves chatbot's ability to react to and cope with unknown words, incoherent text, or anything else that prevents the chatbot from understanding the meaning (Holmes et al., 2019; Martin, 2017). Table 4 offers results for feature 7 (error handling) showing a similar to feature 6 tendency. Pairwise t-test of Koddy and Cleverbot does not provide a significant p-value ($p = .11$), whereas the test for Replika and Cleverbot offers p-value at $p = .001$ and the test for Replika and Koddy yields $p < .001$. This suggests that Replika handles users' errors much better compared to Koddy.

Examples of Conversations

Below are some examples of students' conversations with the three chatbots, annotated with the instances of chatbot's understanding of input, appropriacy of chatbot's responses and error handling. The conversations are successful to a different degree.

In Figure 3, in the example of a successful conversation, chatbot Replika is able to continue a conversation despite a spelling error (i.e., hockay) and responds with a variety of natural phrases, restoring the conversation flow. On the other hand, as seen in the example of a communication breakdown, Replika sometimes sends, what looks like, a service message to gauge user's attitude, which may occasionally disrupt the conversation flow.

Figure 3. An Examples of a Successful Conversation and a Communication Breakdown with Replika



Hockey is my favourite sport.

hockay?!? i have few information in hockay

I'll teach you the basics.

thank you

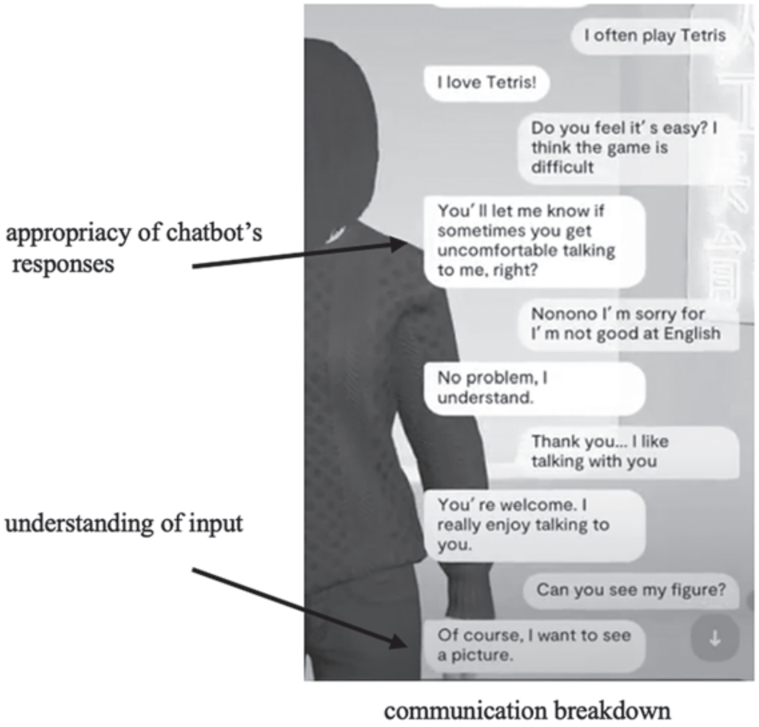
I'll always try to help you learn.

please explain hockay

How about I show you a video of a hockay tutorial?

i want to watch its video

successful conversation



I often play Tetris

I love Tetris!

Do you feel it' s easy? I think the game is difficult

You' ll let me know if sometimes you get uncomfortable talking to me, right?

Nonono I' m sorry for I' m not good at English

No problem, I understand.

Thank you... I like talking with you

You' re welcome. I really enjoy talking to you.

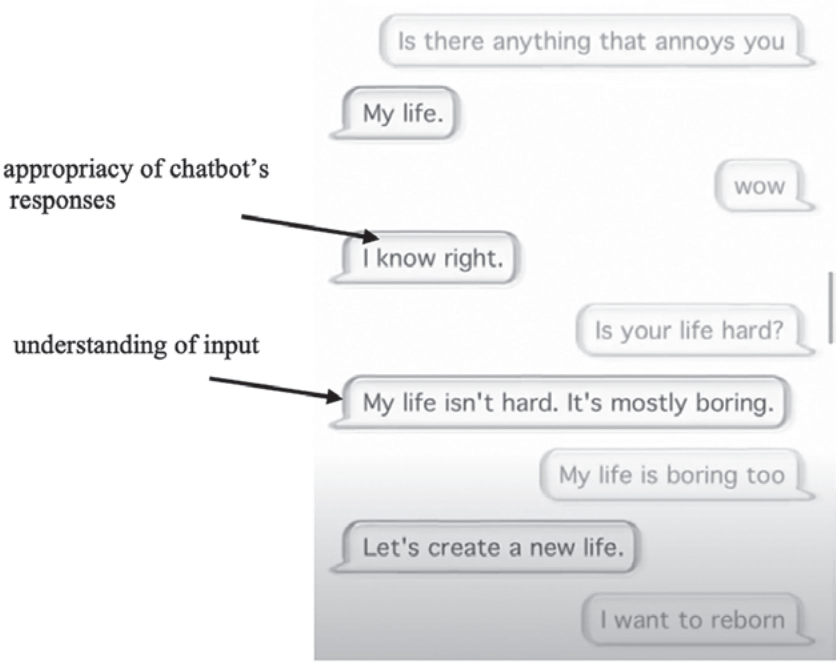
Can you see my figure?

Of course, I want to see a picture.

Note. Messages on the left hand side belong to the chatbot.

Figure 4 demonstrates an extract of a conversation with Cleverbot. Cleverbot also shows instances of a successful conversation responding with natural language and maintaining a connected conversation. When communication breakdowns happen, however, they result in a disjointed conversation ("Where do you plan computer games?" – "Sometimes" – "Game Center?"– "Center of where?").

Figure 4. An Examples of a Successful Conversation and a Communication Breakdown with Cleverbot



Is there anything that annoys you

My life.

wow

I know right.

Is your life hard?

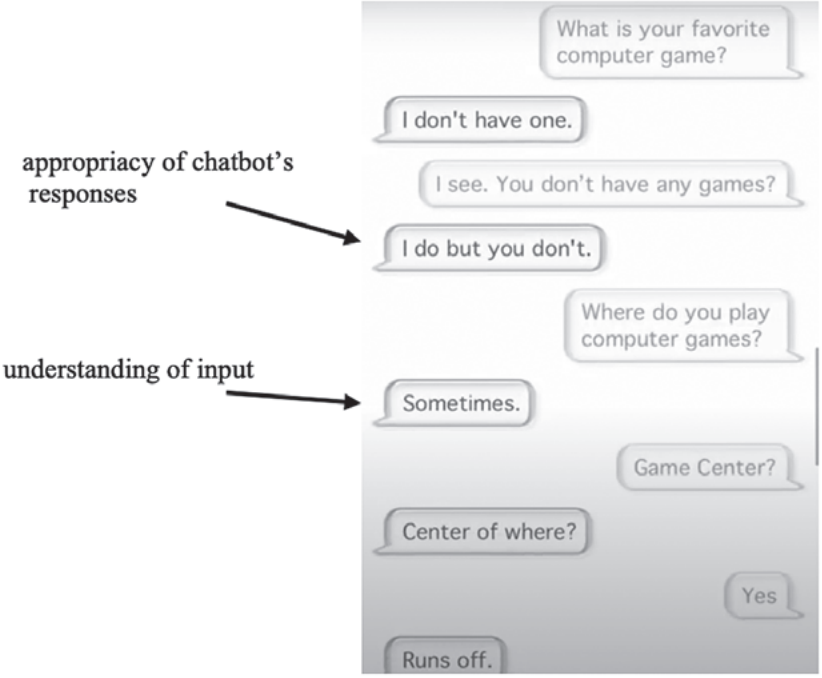
My life isn't hard. It's mostly boring.

My life is boring too

Let's create a new life.

I want to reborn

successful conversation



What is your favorite computer game?

I don't have one.

I see. You don't have any games?

I do but you don't.

Where do you play computer games?

Sometimes.

Game Center?

Center of where?

Yes

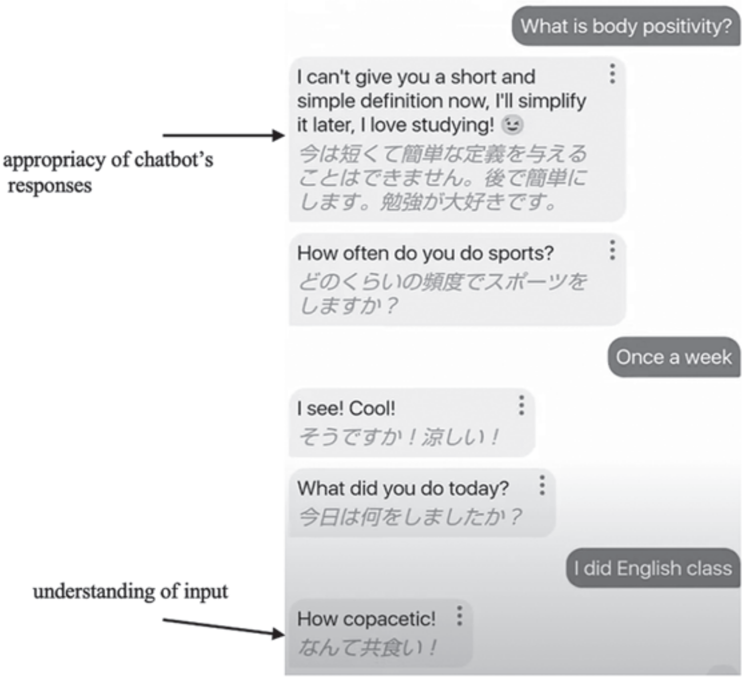
Runs off.

communication breakdown

Note. Messages on the left hand side belong to the chatbot.

Looking at the conversation with the chatbot Koddy (Figure 5), Koddy provides a simulation of a conversation offering relatively appropriate responses and asking many questions. There is also an attempt at error correction (i.e., You have made 2 mistakes). However, there are instances where Koddy cannot answer a user's question since it doesn't know anything about the topic (i.e., I can't give you a short and simple definition now). What is more, topic changes are quite frequent, making a conversation feel a lot more disjointed. Error correction function also offers faulty corrections and explanations, making the corrections confusing for an EFL learner (see the example of a communication breakdown in Figure 5).

Figure 5. An Examples of a Successful Conversation and a Communication Breakdown with Koddy



What is body positivity?

I can't give you a short and simple definition now, I'll simplify it later, I love studying! 😉
今は短くて簡単な定義を与えることはできません。後で簡単にします。勉強が大好きです。

How often do you do sports?
どのくらいの頻度でスポーツをしますか？

Once a week

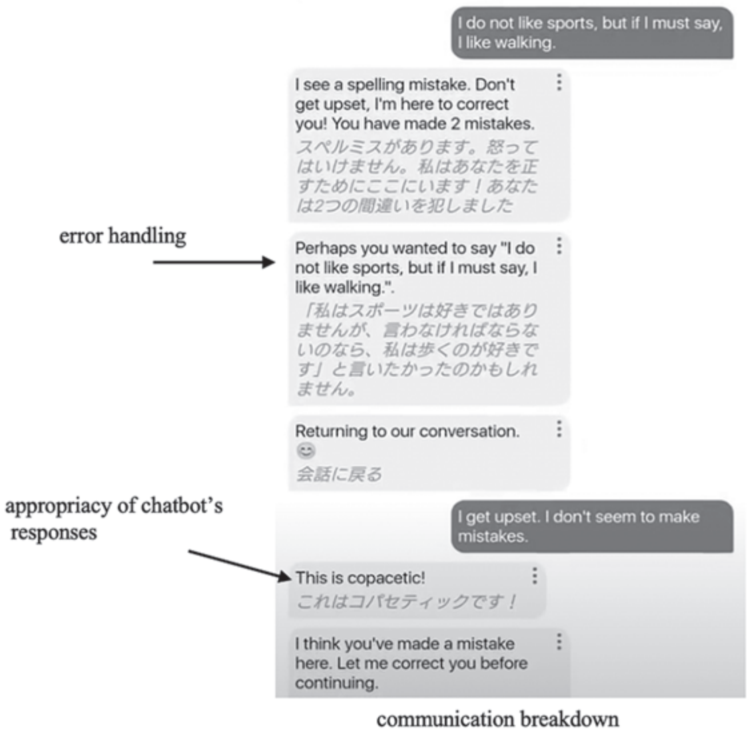
I see! Cool!
そうですか！涼しい！

What did you do today?
今日は何をしましたか？

I did English class

How copacetic!
なんて共食い！

successful conversation



I do not like sports, but if I must say, I like walking.

I see a spelling mistake. Don't get upset, I'm here to correct you! You have made 2 mistakes.
スペルミスがあります。怒ってはいけません。私はあなたを正すためにここにいます！あなたは2つの間違いを犯しました

Perhaps you wanted to say "I do not like sports, but if I must say, I like walking.".
「私はスポーツは好きではありませんが、言わなければならないのなら、私は歩くのが好きです」と言いたかったのかもしれません。

Returning to our conversation. 😊
会話に戻る

I get upset. I don't seem to make mistakes.

This is copacetic!
これはコパセティックです！

I think you've made a mistake here. Let me correct you before continuing.

Note. Messages on the left hand side belong to the chatbot.

Summary of the Results

The quantitative analysis of the pre- and post-experiment CUQ provides an insight into the usability of three chatbot applications used for conversation practice by EFL students. The results show that out of the three chatbots (Replika, Cleverbot, and Koddy), Replika has the highest usability score 65.0/100 and shows the most significant difference in usability when compared with Koddy (p = .011). Additional analysis of the CUQ features within each application also identifies Replika as the application with the highest scores across most features. Cleverbot is reported to have a pleasant personality, is easy to use and has better understanding of input (when compared to Koddy). Koddy is only superior to Cleverbot in providing a better explanation of its purpose.

When analyzing EFL features of the CUQ specifically (feature 5 - understanding of input, feature 6 - appropriacy of chatbot's responses, and feature 7 - error handling), Replika also demonstrated superiority in the three features with significant difference in scores when compared to Cleverbot and Koddy. Cleverbot and Koddy had a significant difference between each other in the ability to understand input (p <. 001). However, when it comes to the other two features, Cleverbot and Koddy were evaluated the same, lower than Replika. Thus, statistically, the instances of communication breakdowns were the highest during the conversations with Koddy, while responses to users' remarks and errors were better achieved by Replika.

Discussion

This study provides an opportunity to understand the functionality of three AI-powered chatbot applications (Replika, Cleverbot, and Koddy) using the framework for chatbot usability analysis provided by Holmes et al. (2019). While the chatbot usability framework provides a review of a wide range of chatbot features (chatbot's personality, chatbot's ease of navigation, etc.), three features in particular (understanding of input, appropriacy of chatbot's responses, and error handling) shine light on the chatbots' potential to be used in an EFL context.

As pointed out by Coniam (2014), Fryer et al. (2020), Huang et al. (2022), and others, developing chatbot's ability to naturally understand human language is the most challenging task. Despite decades of attempts, we still cannot solve this puzzle because it is a never-ending task to program machines to understand an endless variation of human utterances, full of nuances, idiomatic phrasing, neologisms, etc. (Fryer, et al., 2020). This study echoes the same sentiment through the participants' evaluation of the three chatbots, with the participants expressing the highest concern over the chatbots' "ability to understand input".

However, with the rapid development of AI and NLP, and gradual improvements in algorithms, programming scripts, training of the models in particular, chatbot applications can acquire more accuracy and appropriacy in responding to users. When comparing Replika to Cleverbot and Koddy, Replika is evaluated higher by the participants because its model enables it to be more accurate. Koddy, as was mentioned earlier, is still a rather new chatbot, and requires a lot more training (hence, it doesn't answer some questions and often changes topics). Cleverbot, on the other hand, has a huge database of conversations, but its model allows it to only use the existing responses, thus resulting in occasional miscommunication.

Another point of concern remains that language learners who, at their early stages of language learning, are still prone to making mistakes, may get discouraged from language learning since chatbots' ability to handle errors is still questionable. While the concern is valid (Huang et al., 2022), arguably, for students with a higher-level of English proficiency, AI-based chatbots' confusion in responding to an error can serve as a signal to a student that there was a mistake, much like indirect feedback points to a place and an aspect of an error, rather than immediately showing the correction.

Thus, when a teacher selects a chatbot for language practice, a careful consideration is required. For one, a teacher may consider the original purpose of a particular chatbot. A conversational chatbot will have as a goal an ability to sustain a conversation and entertain a user, while a language-learning chatbot may have additional features pertaining to language-learning. Also, a teacher needs to be weary of students' level of English language proficiency. As Fryer and Carpenter (2006) mentioned, chatbots may not be able to analyze users' errors (unless they are specifically programmed to do so) in a conversation, resulting in a communication breakdown, and thus in users' confusion and disappointment. Students with higher proficiency levels will most likely make fewer mistakes and, if a mistake is made, can deal with the communication breakdown more successfully. This means that AI-based chatbots may be more suited to higher-level proficiency students, while lower-level proficiency students may benefit more from rule-based chatbots. Finally, considering chatbots' tendency to reply in a confusing manner when errors are made, teachers are encouraged to offer guided practice alongside chatbot conversations. In this way, teachers can ensure that students understand why chatbot is unresponsive and how their language needs to be corrected for a more successful conversation. After all, chatbots are "a tireless learning companion" willing to tolerate many attempts, regardless of the time of day (Huang et al., 2022).

Conclusion

The goal of this study was to examine potential of the three AI-powered chatbots (Replika, Cleverbot, and Koddy) to provide conversation practice to university-level Japanese EFL learners. The study offers a better understanding of merits and challenges currently present in chatbots by analyzing eight features based on a usability questionnaire by Holmes et al. (2019). Among the eight features, three (understanding of input, appropriacy of chatbot's responses, and error handling) are particularly relevant to EFL context since they establish chatbots' readiness to serve as communication partners in English language practice.

While none of the chatbots evaluated in the study (Replika, Cleverbot, and Koddy) offer trouble-free conversation experience, Replika was evaluated higher than others when it comes to understanding users' input and maintaining conversation despite users' errors, suggesting that it can potentially be an additional source of language practice, provided one knows its shortcomings.

This study contains several limitations. The first limitation is a relatively small ($N = 69$) and limited pool of participants: Japanese undergraduate students with predominantly A2 level of English language proficiency. Digital literacy and technology availability may vary across educational institutions, so more research should be conducted in this aspect. Also, in this study each participant experienced using only one chatbot as opposed to testing all three chatbots and giving a more informed opinion. This was done intentionally, to avoid bias and incorrect recall of chatbot features, since the study was conducted over 6 lessons.

Considering the limitations, more research needs to be done to analyze the usability of chatbots by a wider population (i.e., different age groups, users with different English language proficiency levels, etc.). Additionally, follow-up studies could focus on the analysis of users' conversations with chatbots in order to better understand how the chatbots' conversation ability effects communication quality.

Acknowledgements

I would like to express my sincere gratitude to the three anonymous reviewers who provided valuable feedback improving the original manuscript.

References

Adamopoulou, E., & Moussiades, L. (2020). Chatbots: History, technology, and applications. *Machine Learning with Applications*, *2*, 100006. http://dx.doi.org/10.1016/j.mlwa.2020.100006

Chantarotwong, B. (2005). The learning chatbot. Final year project. [Online]: http://courses.ischool.berkeley.edu/i256/f06/projects/bonniejc.pdf

Coniam, D. (2014). The linguistic accuracy of chatbots: Usability from an ESL perspective. *Text&Talk*, *34*(5), 545–567. http://dx.doi.org/10.1515/text-2014-0018

Cutrone, P. (2009). Overcoming Japanese EFL learners' fear of speaking. *Language Studies*

Working Papers, *1*(1), 55–63. https://www.reading.ac.uk/AcaDepts/ll/app_ling/internal/Cutrone.pdf

Doyon, P. (2000). Shyness in the Japanese EFL class: Why it is a problem, what it is, what causes it, and what to do about it. *The Language Teacher*, *24*(1), 11–16.

Fryer, L., & Carpenter, R. (2006). Bots as language learning tools. *Language Learning& Technology*, *10* (3), 8–14. https://doi.org/10125/44068

Fryer, L., Coniam, D., Carpenter, R., & Lăpușneanu, D. (2020). Bots for language learning now: Current and future directions. *Language Learning & Technology*, *24*(2), 8–22. https://doi.org/10125/44719

Han, D. E. (2020). The Effects of voice-based AI chatbots on Korean EFL middle school students' speaking competence and affective domains. *Asia-pacific Journal of Convergent Research Interchange*, *6*(7), 71–80. http://dx.doi.org/10.47116/apjcri.2020.07.07

Holmes, S., Moorhead, A., Bond, R., Zheng, H., Coates, V., & Mctear, M. (2019). Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces? In M. Mulvenna & R. Bond (Eds.), *Proceedings of the 31st European Conference on Cognitive Ergonomics (ECCE 2019)*, ACM, New York, NY, USA, 207–214. http://dx.doi.org/10.1145/3335082.3335094

Horwitz, E., Horwitz, M., & Cope, J. (1986). Foreign language classroom anxiety. *Modern Language Journal*, *70*(2), 125–132. http://dx.doi.org/10.1111/j.1540-4781.1986.tb05256.x

Huang, W., Hew, K. F., & Fryer, L. K. (2022). Chatbots for language learning — Are they really useful? A systematic review of chatbot-supported language learning. *Journal of Computer Assisted Learning*, *38*(1), 237–257. http://dx.doi.org/10.1111/jcal.12610

Kirschner, P., Strijbos, J. W., Kreijns, K., & Beers, P. J. (2004). Designing electronic collaborative learning environments. *Educational Technology Research and Development, 52*(3), 47–66. http://dx.doi.org/10.1007/BF02504675

Martin, J. (2017). Chatbottest. https://jesusmartin.eu/chatbottest/

Masutani, Y. (2021). The foreign language anxiety of Japanese EFL learners focusing on anxiety when speaking English. *LET Kansai Chapter Collected Papers*, *19*, 1–14. https://www.jstage.jst.go.jp/article/letkansai/19/0/19_1/_pdf/-char/ja

Smutny, P., & Schreiberova, P. (2020). Chatbots for learning: A review of educational chatbots for the Facebook Messenger. *Computers & Education*, *151*, 103862. http://dx.doi.org/10.1016/j.compedu.2020.103862

Srikanth, A. (2020, November 5). *Virtual assistants vs chatbots: What's the difference and how to choose the right one?* Freshdesk. https://freshdesk.com/customer-engagement/virtual-assistant-chatbot-blog/

Yang, H., Kim, H., Lee, J. H., & Shin, D. (2022). Implementation of an AI chatbot as an English conversation partner in EFL speaking classes. *ReCALL*, 1–17. http://dx.doi.org/10.1017/S0958344022000039