EnglishBot: An AI-Powered Conversational System for Second Language Learning

Sherry Ruan*
Stanford University
Stanford, California, USA
ssruan@stanford.edu

Liwei Jiang*
University of Washington
Seattle, Washington, USA
lwjiang@cs.washington.edu

Qianyao Xu*
Tsinghua University
Beijing, China
xuqy17@mail.tsinghua.edu.cn

Glenn M. Davis
Stanford University
Stanford, California, USA
gmdavis@stanford.edu

Zhiyuan Liu
Tsinghua University
Beijing, China
liuzhiyu15@mails.tsinghua.edu.cn

Emma Brunskill
Stanford University
Stanford, California, USA
ebrun@cs.stanford.edu

James A. Landay
Stanford University
Stanford, California, USA
landay@stanford.edu

ABSTRACT

Today, many students learn to speak a foreign language by listening to and repeating pre-recorded materials due to the lack of practice opportunities with human partners. Leveraging recent advancements in AI, Speech, and NLP, we developed EnglishBot, a language learning chatbot that converses with students interactively on college-related topics and provides adaptive feedback. We evaluated EnglishBot against a traditional listen-and-repeat interface with 56 Chinese college students through two six-day user studies under both voluntary and fixed-usage conditions. Students' fluency improved more with EnglishBot as evaluated by the IELTS grading standard for voluntary learning. EnglishBot users also showed higher engagement and voluntarily spent 2.1 times more time interacting with EnglishBot. Our results suggest that conversational interfaces may benefit foreign learners' oral language learning, particularly under casual learning settings.

CCS CONCEPTS

• **Applied computing** → **Education**; • **Human-centered computing** → *Natural language interfaces*; *Empirical studies in HCI*.

KEYWORDS

second language learning, educational chatbots, conversational interface, speech interface

ACM Reference Format:
Sherry Ruan, Liwei Jiang, Qianyao Xu, Glenn M. Davis, Zhiyuan Liu, Emma Brunskill, and James A. Landay. 2021. EnglishBot: An AI-Powered Conversational System for Second Language Learning. In *26th International Conference on Intelligent User Interfaces (IUI '21), April 14–17, 2021, College Station, TX, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3397481.3450648

1 INTRODUCTION

Knowing a foreign language has many benefits and is often a requirement for academic and job opportunities [65]. In China alone, over 300 million people are estimated to be learning English [42]. While many inexpensive self-study resources exist for learning to read and write in a foreign language, opportunities to practice speaking skills are much more limited. Traditional classrooms can be inaccessible, expensive, and may offer only sparse practice opportunities to practice speaking [18]. Intense in-person prep courses are often even more expensive and inaccessible. The lack of opportunity to practice spoken English means that many people may pass written requirements and yet struggle to communicate in a way required by academic training or job opportunities [43, 66]. There is a key need to provide effective education at scale for language learners to practice speaking.

Chatbots are a promising tool to address this. Speech recognition and natural language processesing (NLP) advances have significantly improved chatbot technology, especially in targeted domains [20]. Chatbots can also help simulate the process of talking to another human in a more natural way than synthesized voices and potentially could offer additional benefits such as enhanced distributed cognition and social interaction abilities [35]. New language learners are sometimes shy about practicing their spoken English around other people [51], but chatbots could provide a friendly, non-intimidating setting for spoken language practice. Chatbots may also be more engaging and fun than more traditional spoken language interfaces, such as listen-and-repeat. Perhaps for these reasons, there is emerging commercial interest in foreign

[*] The first three authors contributed equally to this research.

IUI '21, April 14–17, 2021, College Station, TX, USA

ACM ISBN 978-1-4503-8017-1/21/04…$15.00
https://doi.org/10.1145/3397481.3450648

language learning chatbot systems: for example, *English Liulishuo* is used by 150 million people worldwide [40].

Indeed, in other educational settings there is encouraging evidence that chatbots can promote engagement and learning gains [17, 57, 58]. In particular, for factual knowledge learning, prior work [57] found that over a week of optional usage, subjects using a chatbot interface did substantially better on the learning material than those using a classic flashcard app. However, chatbots may also increase the amount of time learners spend with the same material, impacting efficiency of learning.

This suggests that it is important to evaluate and understand the impact of chatbots on spoken language learning effectiveness and efficiency. In this study, we present the first, to our knowledge, experimental study of a standard listen-and-repeat interface versus a chatbot interface (EnglishBot) for spoken language learning. The learning materials are from the standardized English examinations IELTS [61] and TOEFL [54], which cover a range of common scenarios international students could encounter in English-speaking universities. We recruited 56 native Chinese college students who had never worked or studied abroad in English speaking countries for more than three months to participate in two six-day between-subjects studies (one with fixed usage and the other with free usage) comparing EnglishBot against a traditional listen-and-repeat interface. We conducted three assessments to carefully measure users' improvements on memorization of vocabulary, script translation ability, and unrehearsed speaking ability.

Our results show that compared to a traditional listen-and-repeat interface, EnglishBot was more engaging to interact with and promoted more vocabulary learning and speaking fluency and coherence, but did not affect grammatical range and accuracy, lexical resource , or pronunciation. This work makes three contributions: (1) We design a novel conversational interface for spoken language, combining recent advances in natural language processing and speech recognition. (2) We conduct two novel empirical studies evaluating an AI-powered conversational interface against a traditional listen-and-repeat interface for second language speaking under fixed and voluntary usage conditions. (3) In light of the success of chatbots in promoting engagement, vocabulary learning, and speaking fluency, we present design suggestions based on our findings for building the next-generation of conversational interfaces that are both engaging and effective for language learning.

2 RELATED WORK

2.1 Language Learning ITSs

ITSs provide adaptive feedback that depends on the user's response and/or a model of the user's knowledge state, rather than fixed feedback regardless of user input [2, 59]. Prior studies found general ITSs can be as effective as human tutoring [63], and that use of an ITS in a study was associated with greater achievement in comparison with teacher-led large-group instruction, non-ITS computer-based instruction, and textbooks or workbooks [41]. The inclusion of sophisticated AI modules produced learning improvements of 0.3 to 1.0 standard deviations compared with students learning the same content in a classroom [8].

German Tutor [24, 25], later renamed to E-Tutor, was one of the first ITSs developed for foreign language learning. The focus was on building grammatical competence for introductory-level adult learners of German as a foreign language. German Tutor used NLP to parse learner responses to various exercises and provided metalinguistic feedback tailored to the learner's specific errors and perceived proficiency level. Empirical studies of German Tutor/E-Tutor found that students attend to feedback provided by the system [22], and metalinguistic feedback leads to greater learner uptake and self-corrections than simply highlighting mistakes for both grammatical and spelling errors [23, 26].

ITSs for adult Japanese foreign language learners, Nihongo-CALI and BANZAI, have been well-documented in the literature. Through a series of empirical studies, Nagata [47, 48] demonstrated that intelligent metalinguistic feedback provided by the ITSs led to significant improvements on Japanese language tests, as compared to traditional feedback that simply highlighted mistakes. These improvements also held when the ITS was used in a simulated self-study context [49]. The prior work did not use chatbots in their ITSs for language learning.

2.2 Chatbots and Student Learning

While ITSs are able to give more detailed and intelligent feedback to learners than traditional systems, historically such systems assume a constrained set of possible answers from the learner. Chatbots provide the promise of more open-ended interaction between system and learner, in which the learner can take more control over the direction of the conversation, and the system can accommodate a very large range of potential answers. Previous work has shown that chatbots can increase learner engagement [21, 31, 32] and are effective for learning many subjects including factual information [57], math concepts [17, 56], and physics [14–16].

In language education settings, chatbots and conversational agents have been used with learners of all ages. Results have consistently shown that language learners find interacting with intelligent computer agents to be engaging and enjoyable [1, 9, 30, 37, 44–46, 52, 60, 62]. There is some evidence that such systems are more approachable to learners than speaking with human partners, and their use can reduce anxiety about communicating in a foreign language [4, 9, 10, 50, 58].

Although chatbots and conversational agents may be engaging, their actual effectiveness at improving users' foreign language conversation skills has received less study. Selected studies have shown that learners interacting with these systems can produce similar levels of short-term learning in tests of isolated grammar functions, and general speaking proficiency, as interacting with human partners [33, 34, 64]; however, long-term learning has not been studied in detail, and comparisons with traditional non-intelligent systems are limited.

Furthermore, the heightened levels of engagement found with chatbots may represent a short-term novelty effect. In one study, English learners' interest in conducting speaking tasks with a chatbot decreased significantly after the first task, whereas interest in conducting the same tasks with a human partner stayed at the same level through three tasks [11, 12].

Despite limited research on the pedagogical effectiveness of chatbots for foreign language learning, public interest in their development is high. In China, the English learning mobile app *English*

Liulishuo [40] incorporates a conversational agent with adaptive feedback for users to practice dialogues. Given the already high levels of interest in developing chatbots for foreign language learning, it is important to understand whether this increased engagement is also associated with improved foreign language skills in users, as with ITSs, or whether chatbot technology is better suited as a fun activity with limited pedagogical benefit for foreign language learners.

3 SYSTEM

We present the design and implementation details of EnglishBot, a conversational tutor that leverages recent advances of chatbot technology and designed for Chinese-speaking students to practice speaking English. We also describe our baseline listen-and-repeat interface, a traditional medium for practicing English speaking skills, which was used to compare against EnglishBot in our evaluation.

3.1 Learning Materials

Both systems contain the same lessons. We extracted 3 conversations from the listening section of publicly available mock exams for TOEFL (Test of English as a Foreign Language), a standardized English test often required for admission to English-speaking universities[68]. The listening sections of TOEFL exams include rich conversational content on college-related topics, and are of appropriate length and difficulty around which to construct 20-minute lessons. Further, the standardized test aspect allows for easier evaluation using grading rubrics.

Each conversation has two roles; the first to speak ("questioner") begins the conversation with a question, and the second to speak ("responder") provides an answer. As the conversation progresses, both roles ask and answer questions, but we label the roles by their initial line for clarity. The three conversations are about topics relevant to university students: classes and professors, reserving study rooms, and replacing an ID card. In our system, EnglishBot is the first to speak and thus takes the "questioner" role, and the user is the second to speak and thus takes the "responder" role.

3.2 EnglishBot Interface

The EnglishBot system was served as a web application implemented via Python Flask. Figure 1 shows key components of the EnglishBot interface. The EnglishBot interface has two separated and independent panes: the left pane displays learning materials, and the right pane contains a voice chatbot that students converse with to practice spoken English skills.

The learning material pane (left) is comprised of two sections: vocabulary and sentence structures, which were extracted from the contents of the TOEFL mock exam conversations. This pane is used to assist users in completing the conversation fluently and to teach vocabulary and sentence structures in the context of the dialog.

The conversation practice pane (right) serves as the major learning window, where most of the learning actions and interactions take place, as well as all conversations between the user and EnglishBot. The conversation practice pane is comprised of three functional sections. The top section contains dialog bubbles between EnglishBot and the user. EnglishBot's dialog bubbles consist of chatbot utterances, an "audio-play" button (replaying the audio of the English sentences), and a "translation" button, which translates English sentences into Chinese. The user's dialog bubbles contain the user's replies, feedback from EnglishBot, a "reference answer", which the user can click to display the reference answer from the original mock conversation and an "audio play" button that reads the reference answer to the user. The middle section is the "hint" bar containing hints and prompts in the user's native language. The bottom section is the operation panel, where the user records, edits, and sends their replies to EnglishBot's prompts.

We personified EnglishBot by naming it Lily, leveraging the persona effect [38]. Lily plays the role of a teacher as well as a learning companion. All audio content including instructions and the content of the dialog lessons were recorded by a female native English speaker. We chose to use pre-recorded audio instead of relying on a synthesized voice because the human voice sounds more natural and facilitates a relaxed learning environment that is more similar to real-world scenarios. Further, audio content recorded by a native English speaker helps users to compare with their own utterances and correct their pronunciation.

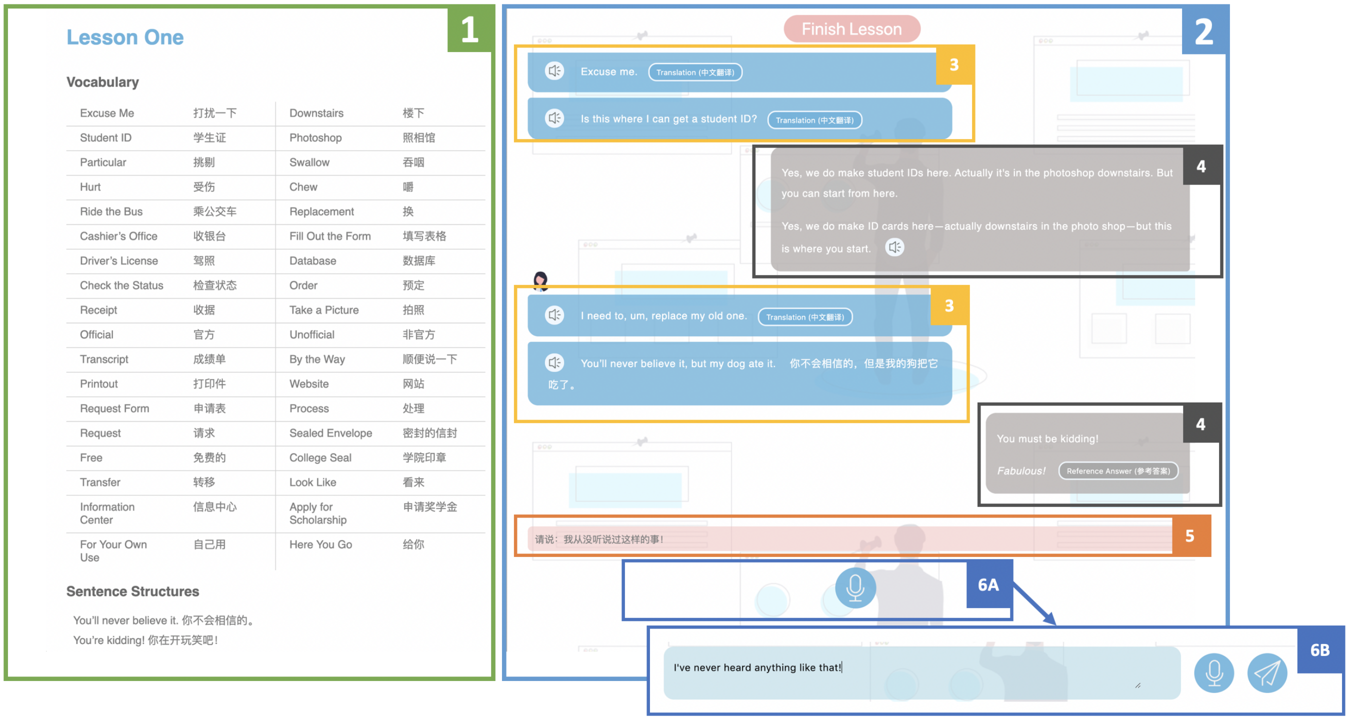
The conversation between the user and EnglishBot begins with a short introduction of how to use the app and an overview of the course structure and content. After the conversation lesson begins, for each round of the conversation, EnglishBot initiates the conversation, and then the user replies based on the hints in Chinese inside the "hint" bar. Audio of EnglishBot's utterances is played automatically. Users can replay the audio and see the Chinese translations of EnglishBot's utterances. After users send a response, the system transcribes the response into a text message format. The user can edit the message by typing into the text field or append new sentences by recording new utterances. After the user approves and sends the response, EnglishBot provides a short feedback sentence (details in the Adaptive Feedback System section). The user can then view the reference answer from the original conversation script, as well as listen to an audio recording of this reference answer. The user can refer to vocabulary and sentence structures in the learning material pane at any time during a learning session.

3.3 Speech Recognition

Users interacted with EnglishBot by speaking. We used Google Chrome's speech recognition API, which required the Chrome browser and a virtual private network (VPN) to access it in mainland China. We used a wired microphone placed close to the user's mouth and asked him or her to speak loudly. We randomly sampled 200 words spoken by users to examine the accuracy of the speech recognition. Out of the 200 words selected, 189 (94.5%) were accurately transcribed. The incorrectly transcribed words can be categorized as 1) long and difficult vocabulary that users pronounced incorrectly (e.g. thermodynamics); 2) near-homophones (e.g. class/clause, more/mall, better/button); 3) indistinguishable liaisons (e.g. "work in"/working).

3.4 Adaptive Feedback System

Because of the importance of adaptive feedback [7, 23, 26, 47, 48, 56], as demonstrated in prior Intelligent Tutoring System (ITS) research, we implemented an adaptive feedback system in EnglishBot.



Lesson One

Vocabulary

Excuse Me	打扰一下	Downstairs	楼下
Student ID	学生证	Photoshop	照相馆
Particular	挑剔	Swallow	吞咽
Hurt	受伤	Chew	嚼
Ride the Bus	乘公交车	Replacement	换
Cashier’s Office	收银台	Fill Out the Form	填写表格
Driver’s License	驾照	Database	数据库
Check the Status	检查状态	Order	预定
Receipt	收据	Take a Picture	拍照
Official	官方	Unofficial	非官方
Transcript	成绩单	By the Way	顺便说一下
Printout	打印件	Website	网站
Request Form	申请表	Process	处理
Request	请求	Sealed Envelope	密封的信封
Free	免费的	College Seal	学院印章
Transfer	转移	Look Like	看来
Information Center	信息中心	Apply for Scholarship	申请奖学金
For Your Own Use	自己用	Here You Go	给你

Sentence Structures

You’ll never believe it. 你不会相信的。
You’re kidding! 你在开玩笑吧!

Excuse me.

Is this where I can get a student ID?

Yes, we do make student IDs here. Actually it's in the photoshop downstairs. But you can start from here.

Yes, we do make ID cards here—actually downstairs in the photo shop—but this is where you start.

I need to, um, replace my old one.

You'll never believe it, but my dog ate it. 你不会相信的，但是我的狗把它吃了。

You must be kidding!

Fabulous!

请说：我从没听说过这样的事!

I've never heard anything like that!

Figure 1: The EnglishBot interface. (1) Learning material pane. (2) Conversation practice pane. (3) EnglishBot’s utterances, with “play-audio” and “translation” buttons. (4) User’s utterances with EnglishBot’s feedback, reference answers, and “play-audio” button. (5) “Hint” bar. (6) Operation panel. (6A) shows the first phase, prompting the user to speak out loud to record their response. (6B) pops up after the user’s voice is recorded and transcribed. The user can click on the “mic” button to continue recording, or edit the text field.

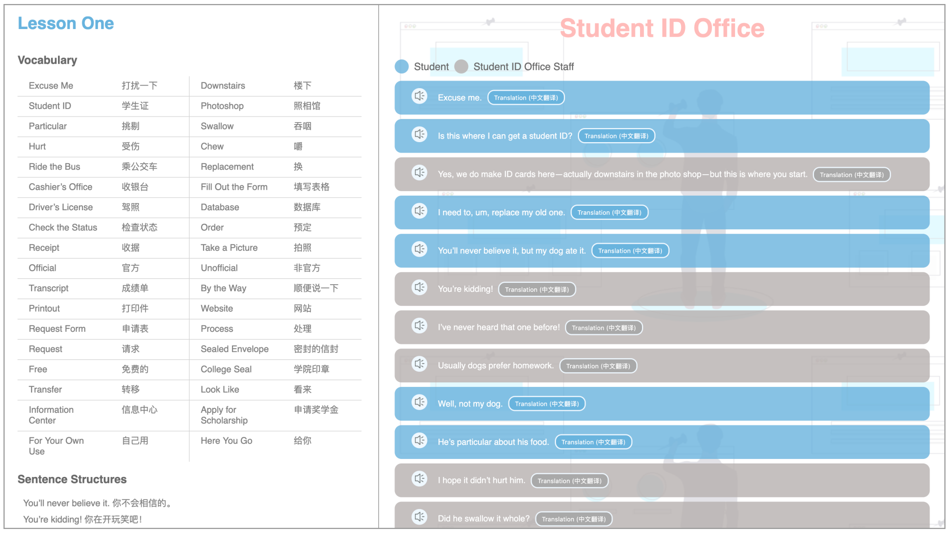
The feedback system compares the user’s transcribed response with the correct response, and scores the user’s response based on two criteria. The first is semantic similarity between the user’s and the correct responses, computed as the cosine similarity between the user and the correct responses using the SIF algorithm [3]. This algorithm represents recent advancements in comparing semantic similarity of sentences and has been used in previous educational chatbot applications [57]. The second is a length score that penalizes short responses to encourage the user to speak more. The length score was computed as the ratio of the length of the user response to the length of the correct answer, capped at 1.

Users of a pilot version of EnglishBot with no feedback completed exercises (n = 264), then rated the quality of feedback from level 1-3 (*1: okay*, *2: good* and *3: excellent*) they would wish to receive based on a comparison between their responses and the correct textbook responses. We excluded negative rating levels, such as *bad* and *terrible*, to balance between constructive and encouraging insights. We then took the similarity score and the length score of these responses as features and fit a logistic regression model using 214 pairs of pilot user responses and labels. We validated the scoring algorithm by testing it on 50 additional pairs of responses and labels. We achieved a mean absolute error (MAE) of 0.52 (about half a level) on the training set and 0.54 on the test set.

The feedback system thus classifies user responses to one of the three performance levels. For each performance level, we manually created a set of 14 encouraging feedback sentences (e.g., *1: You can do this. Try to talk more! 2: Nice work! You are getting better each time. 3: Superb!*). The system randomly selected and displayed a feedback sentence from the corresponding pool, but users received no explicit indication of the performance level determined by the system.

3.5 Traditional Listen-and-Repeat Interface

To compare EnglishBot to more standard approaches, we implemented a traditional-style listen-and-repeat interface based on popular speaking English learning software in China such as *7English* [5], *Little English* [39], and *Lanren English* [69]. As shown in Figure 2, this listen-and-repeat interface was designed to have the same color scheme and layout as EnglishBot, with an identical learning material pane on the left supplementing the main study pane on the right displaying the dialog content. Users can listen to audio recordings and/or see Chinese translations for both “questioner” and “responder” turns, but cannot provide any conversational inputs to the system, which is the major constraint of the popular English learning software systems listed above. By implementing our own listen-and-repeat system, we are able to ensure that the differences between EnglishBot and the comparison interface are minimal, other than the inclusion or exclusion of the conversational elements under study.



Lesson One

Vocabulary

Excuse Me	打扰一下	Downstairs	楼下
Student ID	学生证	Photoshop	照相馆
Particular	挑剔	Swallow	吞咽
Hurt	受伤	Chew	嚼
Ride the Bus	乘公交车	Replacement	换
Cashier’s Office	收银台	Fill Out the Form	填写表格
Driver’s License	驾照	Database	数据库
Check the Status	检查状态	Order	预定
Receipt	收据	Take a Picture	拍照
Official	官方	Unofficial	非官方
Transcript	成绩单	By the Way	顺便说一下
Printout	打印件	Website	网站
Request Form	申请表	Process	处理
Request	请求	Sealed Envelope	密封的信封
Free	免费的	College Seal	学院印章
Transfer	转移	Look Like	看来
Information Center	信息中心	Apply for Scholarship	申请奖学金
For Your Own Use	自己用	Here You Go	给你

Sentence Structures

You’ll never believe it. 你不会相信的。

You’re kidding! 你在开玩笑吧!

Student ID Office

Excuse me.

Is this where I can get a student ID?

Yes, we do make ID cards here—actually downstairs in the photo shop—but this is where you start.

I need to, um, replace my old one.

You’ll never believe it, but my dog ate it.

You’re kidding!

I’ve never heard that one before!

Usually dogs prefer homework.

Well, not my dog.

He’s particular about his food.

I hope it didn’t hurt him.

Did he swallow it whole?

Figure 2: The traditional listen-and-repeat system interface.

3.6 Database

For both systems, we stored all relevant user actions, including conversation logs, button clicks, voice recordings, and corresponding timestamps, in a MongoDB database for managing study progress and further data analysis.

4 USER STUDIES

Similar to previous work [57], we conducted two studies in two different learning scenarios: *fixed usage* and *free usage*, to separate the evaluation of effectiveness from engagement. In Study 1 (fixed usage), users were asked to complete all tasks in one unit each day by following a strict order from beginning to end with no repetition. This study examines the pure efficiency of the EnglishBot system if used in ideal conditions as intended by the researchers and may mimic the use of such a system in a classroom setting, proctored by an English instructor. In Study 2 (free usage), users were given access to one unit each day and allowed to complete as many or as few tasks in the unit as they desired. Users were also free to repeat or skip tasks as they desired. This study examines the effectiveness of EnglishBot in more real-world conditions, and may better simulates the use of such a system outside the classroom, such as during self-study. These studies allowed us to address the following research questions: (1) When interface exposure is held constant, how does the inclusion of chatbot elements affect improvements in users’ English abilities? (2) When users are able to determine their own interface exposure, how does the inclusion of chatbot elements affect improvements in users’ English abilities? (3) How does the inclusion of chatbot elements affect user engagement with the interface?

4.1 Participants

We recruited 62 users (2 dropped out) through social networks, mainly WeChat groups and official accounts. Based on signup order, 4 participated in the pilot study, 28 participated in Study 1, and 28 participated in Study 2. 26 users from study 1 and 19 users from study 2 participated in the three-week delayed followup study. The 56 users (22 men and 34 women) who participated in Studies 1 and 2 had an average age of 23.36 years ($\sigma = 3.92$). Participants came from 22 different universities (19 in China, 3 in Europe; none of the universities used English as the official language of instruction), as well as one high school graduate during a gap year. Over 30 distinct college majors were represented, including computer science, psychology, urban planning, hotel management, medicine, law, engineering, and music. Of the 56 participants in Studies 1 and 2, 45 had never studied or worked in English-speaking countries, and the remaining 11 participants had been abroad for short-term travel purposes only, with average total time in English speaking countries of 1.27 months ($\sigma = 1.62$). Participants were compensated 150 RMB (≈ 21.50 USD) for their six-day commitment.

4.2 Apparatus and Learning Units

All participants were invited to our lab in Beijing for on-site learning sessions to ensure stable Internet and VPN connections and minimize external disturbances. Both EnglishBot and the traditional system were implemented using Python Flask and JavaScript and CSS and were hosted on an Alibaba Cloud server. Two 15-inch MacBook Pros and one iMac were set up for users, and each computer was equipped with a wired earphone with a microphone. Users were randomly assigned to a computer depending on availability. All speaking evaluations were conducted as Zoom conference calls between the participants and examiners. Each study was designed to have six lessons, spanning over six consecutive days. The first three lessons were distinct learning units covering common college environments, and the fourth to sixth lessons repeated the first three units (see Figure 3).

4.3 Evaluation Measures

We included several evaluation measures to both evaluate the progression of users’ English speaking skills and solicit their opinions about the interface and English learning in general.

4.3.1 Spoken English Proficiency. We conducted speaking evaluations before and after using our learning interfaces. The speaking evaluation was comprised of two sections: a free-form conversation and a script-based conversation. In the free-form speaking test, participants had a casual conversation with a native English-speaking tester about topics related to daily college lives from a pre-written pool of nine questions. The free-form speaking test was designed to be two minutes, with 3 to 5 questions depending on the length of participants' responses. In the script-based speaking test, participants were given a script in Chinese and had a English conversation with the same tester according to the pre-written script. All scripts were assembled from Units 1-3 of our systems and all prompts were taken from those units. The script-based speaking test was designed to be three minutes in length, and required participants to translate and speak 12 sentences on the fly without writing the translation down. For both speaking evaluations, we prepared two sets of test materials (question pool for the free-form speaking test and scripts for the scripted speaking test) and counterbalanced the order.

We recruited three testers and two graders to conduct and grade the speaking evaluations, respectively. All three testers were adult native English speakers. All speaking tests were conducted via online Zoom video calls, and all sessions were recorded. The two graders graded all participants' free-form and script-based speaking tests independently according to a grading rubric adapted from the IELTS (International English Language Testing System) speaking test [29]. Graders evaluated the fluency & coherence, lexical resources, grammatical range & accuracy, and pronunciation of each test snippet individually on a scale of 0 to 9. Based on IELTS's grading guidelines, these four scores were averaged to yield a final speaking score on a scale of 0 to 9. Inter-rater reliability between the two graders was measured to be 0.722 using Krippendorff's alpha [19, 36]. The final scores for each test snippet are the average scores of the two graders.

4.3.2 English Vocabulary Acquisition. We administered the same vocabulary test on the first and last day of using the interface, and also three weeks after the last day of usage to measure retention. In this test, users were given a sheet of 36 words in Chinese and were asked to provide spoken English translations. All words on the vocabulary test were taken from the featured key words of each unit. A random subset of 18 words from each role (questioner, responder) were selected.

4.3.3 Foreign Language Anxiety. Previous research has found that anxiety about using a foreign language is associated with worse oral exam performance [27]. In this study, we measure users' foreign language anxiety with the Second Language Speaking Anxiety Scale (SLSAS) [67]. This scale builds on the well-known Foreign Language Classroom Anxiety Scale (FLCAS) [28] in order to assess foreign language anxiety both in and out of classroom contexts. Participants were asked to fill out the SLSAS on the first and last days of using the learning system.

4.3.4 Engagement Metrics. After the last day of using the system, we administered the User Engagement Scale Short Form [53], a popular survey for measuring people's engagement with software. It contains 12 questions and takes about 15 minutes to finish.

4.3.5 English Language Learning Experiences and Motivations. We conducted a survey to understand participants' opinions about oral English study prior to their participation in our studies. We found that our participants' motivations for studying English varied drastically, but were mostly driven by compulsory external factors such as school curricula and high-stakes examinations (89.3% of participants) rather than voluntary interest in self-advancement or international communication (10.5%). When comparing the four basic language skills (reading, writing, speaking, and listening), the largest number of participants rated speaking as the most important skill (46.4%), the most useful skill (60.7%), and the most difficult one to learn (48.2%).

We also asked participants the biggest challenges they face when speaking English. Participants reported being self-conscious about their pronunciation (21.4%) and about making mistakes in front of others (14.3%), and noted that finding people to practice oral English with was difficult (19.6%). These survey findings reinforce the need for an effective and engaging learning system that can help students improve oral English skills without fear of embarrassment.

4.4 Study Procedure

We ran two between-subject studies. Both studies shared the same study flow as shown in Figure 3. Each study lasted for six consecutive days. During the first day, participants took the pre-survey containing the foreign language anxiety measure, followed by the speaking test and vocabulary test. Then participants were randomly assigned to an interface (EnglishBot or traditional-style) and used it for six days in a row. During the last day, participants took the post-survey containing the same language anxiety measure, the engagement scale, the speaking test, and the same vocabulary test with the same tester.

We first ran a **fixed usage** between-subjects study where the number of practices was fixed. We then ran a second **free usage** between-subjects study where users used the system voluntarily. We aimed to answer how engaged people are with both systems with the first study, and how people's speaking English might change when practicing the same number of sentences.

4.4.1 Between-Subjects Study 1: Fixed Usage. In Study 1, users were required to use the learning system exactly as prescribed. For EnglishBot, users were asked to take each lesson exactly once. Within each lesson, they were asked to listen to the questioner's audio recordings once, which were prompted by EnglishBot automatically. Users then spoke according to the hint given by the system, and recorded their answers in response. After they sent out their answers and received feedback from the system, they were instructed to click on the "reference answer" button to read the reference answer text and listen to its audio once. For the traditional interface, users were asked to listen to all audio recordings exactly once in the order they appear. For the questioner's content, they were asked to not repeat it after listening to the audio, and for the responder's content, they were required to repeat it once after listening to the audio.

4.4.2 Between-Subjects Study 2: Free Usage. In Study 2, users were asked to use the systems voluntarily to practice oral English for six days. For both systems, users were given access to one unit each

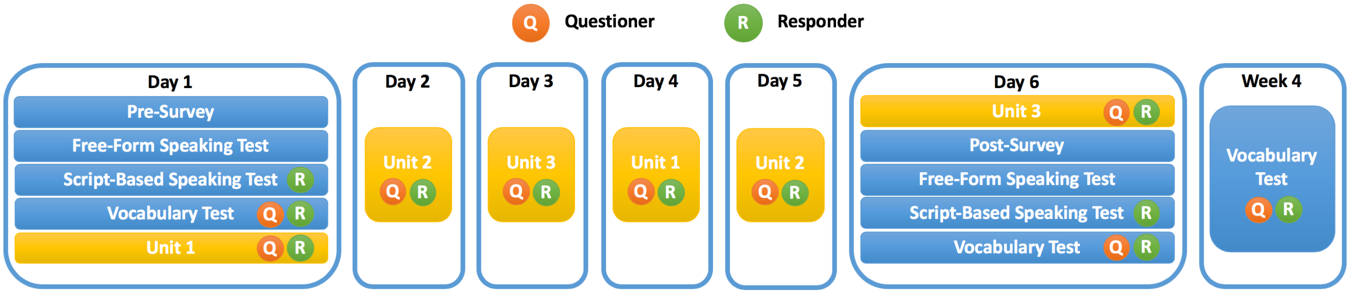


Figure 3: User study procedure.

day, as in Figure 3. There were no specific requirements on usage time and users were free to explore any functionality within that unit, as many times as they wished.

5 EVALUATION RESULTS

In this section, we present the evaluation results from the two studies described above. We performed Shapiro-Wilk normality tests before running all the t-tests.

5.1 Study 1: Fixed Usage

5.1.1 Usage Time, Engagement, and Anxiety Change. In the fixed usage condition, the traditional interface users spent a total of 41.08 ($\sigma = 11.13$) minutes using the app, and the EnglishBot users spent in total 105.70 ($\sigma = 29.63$) minutes (Figure 5). Thus, EnglishBot users spent on average 2.6 times as long on the same learning materials, a statistically significant difference with a large effect size (Cohen's d [6] = 2.89), shown by a two-sample t-test ($t_{26} = 7.6, p < .0001$).

Figure 4 shows users' self-reported user engagement score evaluated with the UES Short Form [53] after the six-day learning period. EnglishBot users' engagement rating (4.18) was higher than that of the traditional system user sessions (3.94), but the difference was not statistically significant, as revealed by a two-sample t-test ($t_{26} = 1.1, p > .5$).

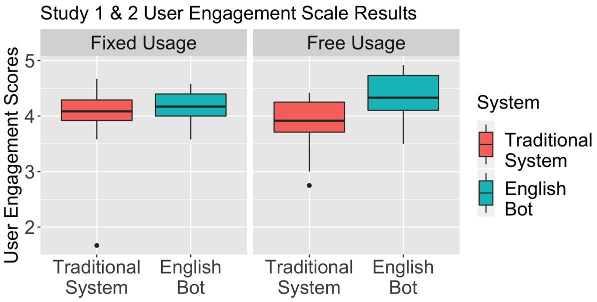


Figure 4: Self-reported user engagement results, Study 1 & 2.

After six days, the traditional system users' foreign language speaking anxiety evaluated with SLSAS [67] decreased by 0.12 ($\sigma = 0.59$) and EnglishBot users' anxiety decreased by 0.03 ($\sigma = 0.34$), but the changes were not significant (both $ps > .05$), shown by one-sample t-tests. Further, two-sample t-tests demonstrated that the two learning systems did not significantly differ in terms of foreign language speaking anxiety change ($t_{26} = 0.5, p > .05$).

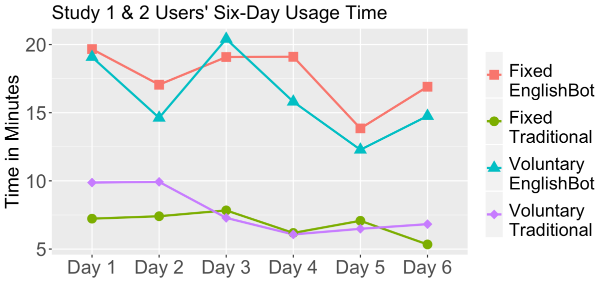
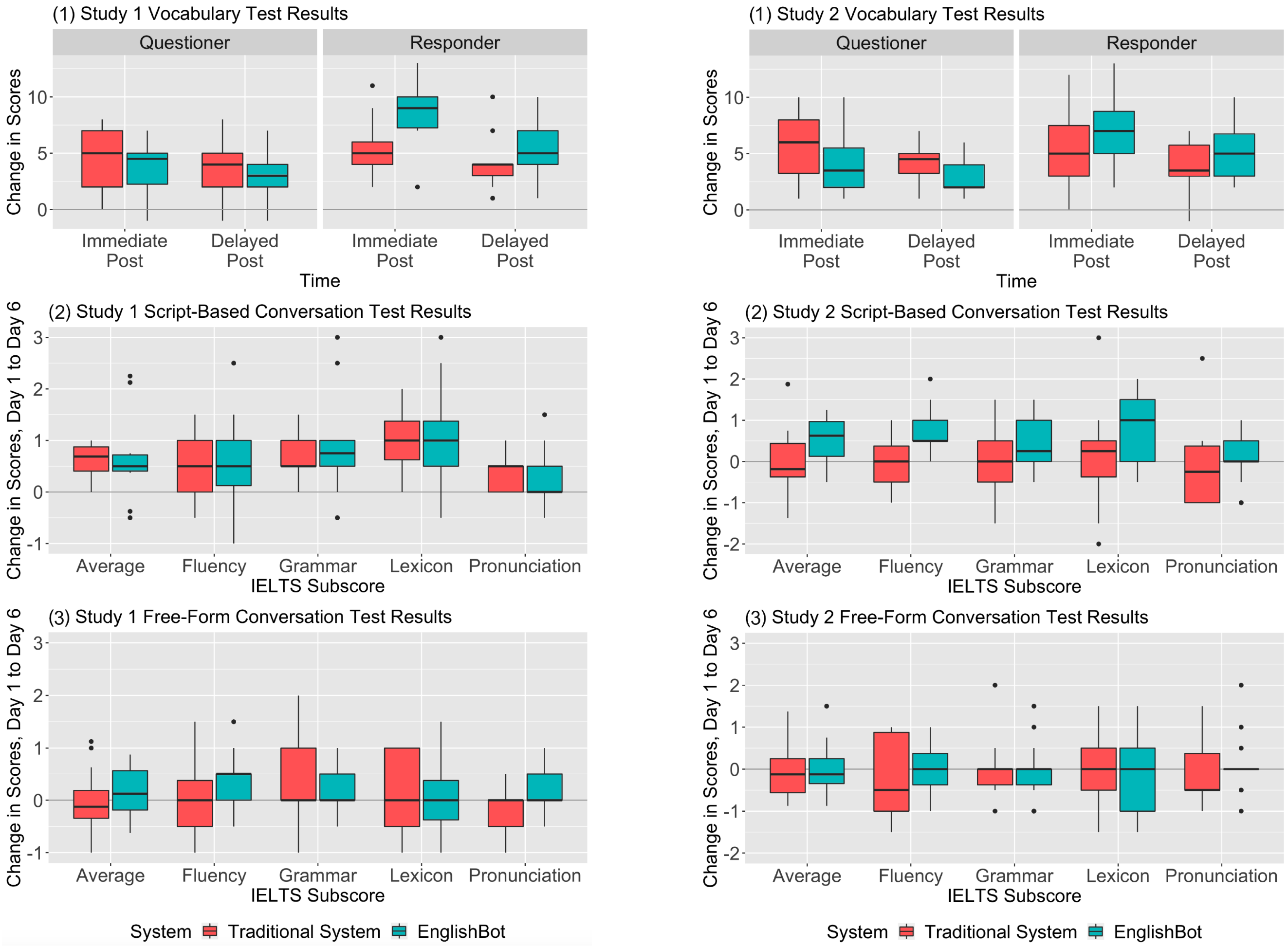


Figure 5: User's six-day usage time with two systems under two study conditions.

5.1.2 Vocabulary Test Results. Results of an immediate and a delayed vocabulary test are presented in Figure 6a (1). Since users interacted with learning materials in the questioner's and the responder's role differently, we separated the analysis of vocabulary test results of these two roles and corrected them with Bonferroni's procedure. One-sample t-tests showed that, immediately after the six-day learning period, traditional system users had learned 4.4 new words ($\sigma = 2.6$) on the questioner's side ($p < .001$, Cohen's d = 2.35) and 5.4 new words ($\sigma = 2.6$) on the responder's side ($p < .0001$, Cohen's d = 2.13). Taking into account users' initial vocabulary test performance, users learned 62.7% ($\sigma = 22.5\%$) of the new words in the questioner's role and 81.3% ($\sigma = 21.1\%$) of the new words in the responder's role.

EnglishBot users learned 3.8 new words ($\sigma = 2.3$) on the questioner's side ($p < .001$, Cohen's d = 2.02) and 8.6 new words ($\sigma = 2.6$) on the responder's side ($p < .0001$, Cohen's d = 4.31), equivalent to 44.0% ($\sigma = 21.8\%$) of new words on the questioner's side and 88.1% ($\sigma = 16.9\%$) of new words on the responder's side. When comparing users' vocabulary improvements with the two systems, two-sample t-tests showed that EnglishBot users learned significantly more words than the traditional system users on the responder's role ($t_{26} = 3.4, p < .005$, Cohen's d = 1.27) and a similar amount of words on the questioner's role ($t_{26} = 0.7, p > .05$).

Figure 6a (1) also depicts three-week delayed vocabulary test results. Although they decreased over time, users' vocabulary improvements were still significant with either system and on either role ($p < .001$ for traditional system on the questioner's role,



(a) Study 1 Improvements

(b) Study 2 Improvements

Figure 6: Improvements in vocabulary, script-based speaking, and free-form speaking test scores from pre-test to post-test in (a) study 1 and (b) study 2.

$p < .001$ on the responder's role, $p < .01$ for EnglishBot on the questioner's role, and $p < .0001$ on the responder's role). However, the difference between the two systems was not statistically significant for either role ($ps > .05$).

5.1.3 Speaking Test Results. Figure 6a (2) and Table 1a show changes in script-based speaking test scores from Day 1 to 6. We conducted a series of 10 one-sample t-tests to determine whether significant (non-zero) improvements were achieved for any subscores (fluency, grammar, lexicon, pronunciation, overall average) by either system. Applying the Bonferroni correction for multiple tests, we find that on the script-based speaking test, EnglishBot produced a significant improvement in lexicon subscores ($p < .05$) with a large effect size (Cohen's $d = 0.93$), and the traditional system produced a significant improvement in all the subscores ($p < .05$, Cohen's $d = 0.47$ for fluency, $p < .0001$, Cohen's $d = 0.67$ for grammar, $p < .0001$, Cohen's $d = 0.86$ for lexicon, and $p < .05$, Cohen's $d = 0.42$ for pronunciation) and the overall average speaking test scores ($p < .0001$, Cohen's $d = 0.67$). On the free-form speaking test, neither system produced a significant improvement in any of the subscores or the overall scores, as shown in Figure 6a (3) and Table 1a.

Two-sample t-tests revealed that the systems did not significantly differ in terms of improvements on any of the four speaking test subscores or on the overall average scores for either script-based or free-form speaking tests ($ps > .05$).

5.2 Study 2: Free Usage

5.2.1 Usage Time, Engagement, and Anxiety Change. In this condition, users spent 97.0 ($\sigma = 19.9$) minutes learning with the EnglishBot system and 46.5 ($\sigma = 23.9$) minutes with the traditional system. Of their own volition, EnglishBot users spent 2.1 times more time than the traditional system users, and the difference was statistically significant ($t_{26} = 6.1, p < .0001$) with a large effect size

(a) Study 1 Improvement Results

Subscore	Interface	Script Imp.	Free Imp.
Fluency	Traditional	0.500*	-0.036
Fluency	EnglishBot	0.607	0.214
Grammar	Traditional	0.643***	0.214
Grammar	EnglishBot	0.821	0.250
Lexicon	Traditional	1.000***	0.036
Lexicon	EnglishBot	0.964*	0.107
Pron.	Traditional	0.393*	-0.214
Pron.	EnglishBot	0.179	0.214
Average	Traditional	0.634***	0.000
Average	EnglishBot	0.643	0.196

(b) Study 2 Improvement Results

Subscore	Interface	Script Imp.	Free Imp.
Fluency	Traditional	-0.036	-0.214
Fluency	EnglishBot	0.679**	0.000
Grammar	Traditional	0.000	0.036
Grammar	EnglishBot	0.429	-0.036
Lexicon	Traditional	0.107	0.071
Lexicon	EnglishBot	0.786*	-0.107
Pron.	Traditional	-0.143	-0.143
Pron.	EnglishBot	0.143	0.107
Average	Traditional	-0.018	-0.063
Average	EnglishBot	0.509*	-0.009

Table 1: Improvements in script-based and free-form conversation test scores from pre-test to post-test, Study 1 and 2. * indicates a significant non-zero improvement with corrected Bonferroni p < .05, ** with corrected p < .01, * with corrected p < .0001.**

(Cohen's d = 2.30). The learning time over the six days are shown in Figure 5.

Users' average engagement rating was 4.36 (σ = 0.43) for EnglishBot and 3.87 (σ = 0.51) for the traditional system, as shown in Figure 4. A t-test showed that users rated EnglishBot as significantly more engaging than the traditional system ($t_{26} = 2.7, p < .05$, Cohen's d = 1.03), suggesting that users found EnglishBot more engaging to use in the casual learning setting.

Traditional system users' speaking English anxiety increased by 0.065 (σ = 0.381) and EnglishBot users' speaking English anxiety increased by 0.006 (σ = 0.414), but neither of the changes was significant (ps> .05), nor was there any difference between the two systems ($t_{26} = 0.4, p > .05$).

5.2.2 Vocabulary Test Results. Figure 6b (1) shows users' vocabulary score improvements. On the questioner's end, traditional system users learned 5.6 (σ = 2.8) new words, accounting for 71.6% (σ = 22.7%) of previously unknown words in the pre-test, and the improvement was significant ($p < .0001$). EnglishBot users learned 4.0 (σ = 2.6) new words, 59.0% (σ = 28.3%) of previously unknown words in the pre-test, and the improvement was also significant ($p < .001$). On the responder's side, traditional system users memorized 5.0 (σ = 3.5) more new words ($p < .001$), 74.8% (σ = 21.4%) improvement comparing to the initial vocabulary test performance, and EnglishBot users memorized 6.9 (σ = 2.9) more new words ($p < .0001$), equivalent to 93.7% (11.3%) improvement against initial performance. There was no difference on vocabulary improvements on the questioner's side ($t_{26} = 1.5, p > .05$) or the responder's side ($t_{26} = 1.6, p > .05$) between the two systems.

For the three-week delayed vocabulary test, traditional system users retained 2.3 (σ = 6.6) words on the questioner's side ($p <$ .0001), and 2.0 (σ = 6.7) on the responder's side ($p < .01$). EnglishBot users retained 2.9 (σ = 1.6) words ($p < .01$) on the questioner's side, and 5.4 (σ = 2.9) on the responder's side ($p < .01$). We did not see any difference between the two systems on delayed vocabulary test results on the questioner's side ($t_{17} = 0.1, p > .05$) or the responder's side ($t_{17} = 1.8, p > .05$).

5.2.3 Speaking Test Results. Figure 6b (2, 3) and Table 1b show the changes in speaking test scores from Day 1 to Day 6. As shown in Table 1b, one-sample t-tests indicated that on the script-based speaking test, EnglishBot produced significant (non-zero) improvements in overall average speaking test scores ($p < .05$, Cohen's d = 0.64) and in the fluency ($p < .01$, Cohen's d = 0.84) and lexicon subscore (p < .05, Cohen's d = 0.83). Traditional system did not produce any significant improvements (ps> .05). Neither system produced significant improvements on any of the subscores or the overall scores on the free-form speaking test (ps> .05).

Two-sample t-tests revealed that EnglishBot produced significantly better improvements in the fluency subscore ($t_{26} = 3.1, p <$ .05, Cohen's d 1.18) on the script-based speaking test, and not in the other speaking test subscores or the overall average scores (ps > .05). Further, after normalizing users' improvements on fluency by total time spent, the difference between the two systems was not significant any more ($p > .05$). There was no difference on users' free-form speaking test improvements between the two systems ($p > .05$).

6 DISCUSSION

We present key study insights, experimental limitations, along with design suggestions for building future intelligent conversational interface for second language learning based on our findings.

6.1 Voluntary Usage of EnglishBot Promotes Engagement

We found that when usage restrictions were lifted and users were free to engage with the learning interfaces however they wanted, users spent more time interacting with EnglishBot, and they also reported finding it more engaging than the traditional interface. That conversational interfaces promote engagement is consistent with previous studies of educational chatbots [4, 13, 56, 58].

At the same time, users did not find EnglishBot more engaging in the fixed usage setting. This difference from the fixed usage setting can be perhaps attributed to familiarity: traditional-style listen-and-repeat systems are sometimes used as part of classroom curricula, where their usage is enforced by instructors. When users are voluntarily interacting with these traditional systems, there may still be associations with unpleasant classroom studying experiences that are not present with a more novel chatbot-based learning interface.

6.2 Limitations on the Effectiveness Results

EnglishBot users in the free usage condition significantly improved their scores in the script-based conversation test, including the fluency and lexicon subscores and overall scores, whereas traditional system users showed no improvements.

The results suggest the inclusion of chatbot elements can increase the benefits of a system for learning English as a foreign language. EnglishBot was found to be more engaging, and it led to some minor improvements in speaking skills that were not found with the traditional system. However, it is crucial to note that the script-based conversation test focuses more on rote memorization of previous spoken lines, and neither system in the two studies led to improvements in the free-form conversation test, which more closely represents English proficiency in real-world scenarios. This may be due to the short-term learning period allowed for the users. Common sense and expert knowledge demonstrate that mastering a second language is a complicated process that happens through and over time [55]. Therefore, one important direction for future work is a longitudinal study to examine the effectiveness of using conversational agents to learn a foreign language.

6.3 Conversational Interface Design Implications

Overall, the conversational interface provides a more engaging and immersive learning experience than the listen-and-repeat interface. Based on the participants' qualitative response, we present the following design suggestions for building more effective and engaging intelligent conversational interfaces for language learning.

Adaptive Feedback. The EnglishBot feature that was referred to as "useful" by the greatest number of people ($n = 11$) is real-time adaptive feedback on their speaking during the conversation. It is crucial to provide positive feedback to learners to boost their learning confidence and encourage them to speak more. For future improvement, our users suggested more detailed feedback based on both content (i.e., word choice, grammar, sentence structure, etc.) and pronunciation.

Audio and Speech Recognition. Employing high quality pre-recorded audio for the conversational lessons is especially helpful for learners to make self-correction of their speaking tones, speed, and pronunciation ($n = 10$). Recording and replaying the learners' responses and comparing them to pre-recorded answers helps students to self-reflect for faster improvement.

Conversation Style. Besides practicing structured content, some of our users ($n = 3$) find that having random conversations and chitchat will substantially enhance their engagement and speaking skills in the long run. Additionally, instead of reading hints in text form, incorporating them into the conversation verbally can make practicing more effective.

Incorporating these design suggestions into our sysetm as well as running longitudinal studies to examine students' speaking ability improvement more thoroughly remains as critical future work.

7 CONCLUSION

Seeing successful applications of chatbots to other learning subjects, we built EnglishBot, an interactive chatbot that helps students practice their foreign language conversation skills. Evaluating EnglishBot against a traditional listen-and-repeat interface with 56 students, we found that chatbots offered a more engaging interface. We also observed small improvements on students' vocabulary learning, speaking fluency and coherence. We conclude with design suggestions on further improving the engagement and effectiveness of intelligent conversational interfaces for language learning.

ACKNOWLEDGMENTS

We thank Lily Zhou for recording English audio, Catherine Wang and Kyle Nguyen for making the complementary video.

REFERENCES

[1] James N. Anderson, Nancie Davidson, Hazel Morton, and Mervyn A. Jack. 2008. Language learning with interactive virtual agent scenarios and speech recognition: Lessons learned. *Computer Animation and Virtual Worlds* 19 (2008), 605–619.

[2] John R. Anderson, C. Franklin Boyle, and Brian J. Reiser. 1985. Intelligent Tutoring Systems. *Science* 228, 4698 (1985), 456–462.

[3] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A Simple but Tough-to-Beat Baseline for Sentence Embeddings. *International Conference on Learning Representations* (2017).

[4] Emmanuel Ayedoun, Yuki Hayashi, and Kazuhisa Seta. 2019. Adding communicative and affective strategies to an embodied conversational agent to enhance second language learners' willingness to communicate. *International Journal of Artificial Intelligence in Education* 29 (2019), 29–57.

[5] Chongqing Mizao Technology Co. 2020. 7English. http://www.7english.cn/index.html. Accessed: 2020-10-05.

[6] Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences.* Routledge.

[7] Albert Corbett. 2001. Cognitive computer tutors: Solving the two-sigma problem. In *International Conference on User Modeling*. Springer, 137–147.

[8] Albert Corbett, John Anderson, Art Graesser, Ken Koedinger, and Kurt VanLehn. 1999. Third generation computer tutors: Learn from or ignore human tutors?. In *Conference on Human Factors in Computing Systems - Proceedings*. 85–86.

[9] Luke Fryer and Rollo Carpenter. 2006. Emerging technologies: Bots as language learning tools. *Language Learning & Technology* 10, 3 (2006), 8–14.

[10] Luke Fryer and Kaori Nakao. 2009. Shared identities: Our interweaving threads. In *JALT2008 conference proceedings*, Alan Stoke (Ed.). JALT, Tokyo, 259–274.

[11] Luke K. Fryer, Mary Ainley, Andrew Thompson, Aaron Gibson, and Zelinda Sherlock. 2017. Stimulating and sustaining interest in a language course: An experimental comparison of Chatbot and Human task partners. *Computers in Human Behavior* 75 (2017), 461–468. https://doi.org/10.1016/j.chb.2017.05.045

[12] Luke K. Fryer, Kaori Nakao, and Andrew Thompson. 2019. Chatbot learning partners: Connecting learning experiences, interest and competence. *Computers in Human Behavior* 93 (2019), 279–289. https://doi.org/10.1016/j.chb.2018.12.023

[13] Yoshiko Goda, Masanori Yamada, Hideya Matsukawa, Kojiro Hata, and Seisuke Yasunami. 2014. Conversation with a Chatbot before an Online EFL Group Discussion and the Effects on Critical Thinking. *The Journal of Information and Systems in Education* 13, 1 (2014), 1–7. https://doi.org/10.12937/ejsise.13.1

[14] Arthur C Graesser, Patrick Chipman, Brian C Haynes, and Andrew Olney. 2005. AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions on Education* 48, 4 (2005), 612–618.

[15] Arthur C Graesser, Shulan Lu, George Tanner Jackson, Heather Hite Mitchell, Mathew Ventura, Andrew Olney, and Max M Louwerse. 2004. AutoTutor: A tutor with dialogue in natural language. *Behavior Research Methods, Instruments, & Computers* 36, 2 (2004), 180–192.

[16] Arthur C Graesser, Katja Wiemer-Hastings, Peter Wiemer-Hastings, Roger Kreuz, Tutoring Research Group, et al. 1999. AutoTutor: A simulation of a human tutor. *Cognitive Systems Research* 1, 1 (1999), 35–51.

[17] Joshua Grossman, Zhiyuan Lin, Hao Sheng, Johnny Tian-Zheng Wei, Joseph Jay Williams, and Sharad Goel. 2019. MathBot: Transforming Online Resources for Learning Math into Conversational Interactions. *AAAI 2019 Story-Enabled Intelligence* (2019).

[18] Joel Heng Hartse, Jiang Dong, and Andy Curtis. 2015. *Perspectives on teaching English at colleges and universities in China: English Language Teaching and Learning at the National Level in China.* TESOL Press, 5–13.

[19] Andrew F Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication methods and measures* 1, 1 (2007), 77–89.

[20] Marti A. Hearst. 2015. Can Natural Language Processing Become Natural Language Coaching?. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Beijing, China, 1245–1252. https://doi.org/10.3115/v1/P15-1120

[21] Neil T Heffernan and Kenneth R Koedinger. 2002. An intelligent tutoring system incorporating a model of an experienced human tutor. In *International Conference on Intelligent Tutoring Systems*. Springer, 596–608.

[22] Trude Heift. 2001. Error-specific and individualised feedback in a Web-based language tutoring system: Do they read it? *ReCALL* 13, 1 (2001), 99–109.

[23] Trude Heift. 2004. Corrective feedback and learner uptake in CALL. *ReCALL* 16, 2 (2004), 416–431. https://doi.org/10.1017/S0958344004001120

[24] Trude Heift and Devlan Nicholson. 2000. Theoretical and practical considerations for web-based intelligent language tutoring systems. In *ITS 2000: Intelligent tutoring systems: 5th international conference*, G. Gauthier, C. Frasson, and K. VanLehn (Eds.). Springer, 354–362.

[25] Trude Heift and Devlan Nicholson. 2001. Web delivery of adaptive and interactive language tutoring. *International Journal of Artificial Intelligence in Education* 12 (2001).

[26] Trude Heift and Anne Rimrott. 2008. Learner responses to corrective feedback for spelling errors in CALL. *System* 36 (2008), 196–213. https://doi.org/10.1016/j.system.2007.09.007

[27] Elaine Hewitt and Jean Stephenson. 2012. Foreign language anxiety and oral exam performance: A replication of Phillips's MLJ study. *The Modern Language Journal* 96, 2 (2012), 170–189. https://doi.org/10.1111/j.1540-4781.2011.01174.x

[28] Elaine K. Horwitz, Michael B. Horwitz, and Joann Cope. 1986. Foreign Language Classroom Anxiety. *The Modern Language Journal* 70, 2 (1986), 125–132.

[29] IELTS. [n.d.]. IELTS Scoring In Detail. https://www.ielts.org/en-us/ielts-for-organisations/ielts-scoring-in-detail. Accessed: 2020-10-05.

[30] Jiyou Jia and Weichao Chen. 2008. Motivate the learners to practice English through playing with chatbot CSIEC. In *Edutainment 2008: Technologies for E-learning and digital entertainment (Lecture Notes in Computer Science 5093)*, Zhigeng Pan, Xiaopeng Zhang, Abdennour El Rhalibi, Woontack Woo, and Yi Li (Eds.). Springer, 180–191. https://doi.org/10.1007/978-3-540-69736-7_20

[31] W Lewis Johnson and James C Lester. 2018. Pedagogical Agents: Back to the Future. *AI Magazine* 39, 2 (2018).

[32] Greg Jones and Scott Warren. 2009. The Time Factor: Leveraging Intelligent Agents and Directed Narratives in Online Learning. *Innovate: Journal of Online Education* 5, 2 (2009), 2.

[33] Na-Young Kim. 2016. Effects of voice chat on EFL learners' speaking ability according to proficiency levels. *Multimedia-Assisted Language Learning* 19, 4 (2016), 63–88.

[34] Na-Young Kim. 2019. A study on the use of artificial intelligence chatbots for improving English grammar skills. *Journal of Digital Convergence* 17, 8 (2019), 37–46. https://doi.org/10.14400/JDC.2019.17.8.037

[35] Yanghee Kim and Amy L Baylor. 2006. A social-cognitive framework for pedagogical agents as learning companions. *Educational technology research and development* 54, 6 (2006), 569–596.

[36] Klaus Krippendorff. 2004. Content analysis: An introduction to its methodology (2nd ed.). *Sage publications* (2004), 241.

[37] Sungjin Lee, Hyungjong Noh, Jonghoon Lee, Kyusong Lee, Gary Geunbae Lee, Seongdae Sagong, and Munsang Kim. 2011. On the effectiveness of robot-assisted language learning. *ReCALL* 23, 1 (2011), 25–58. https://doi.org/10.1017/S0958344010000273

[38] James C Lester, Sharolyn A Converse, Susan E Kahler, S Todd Barlow, Brian A Stone, and Ravinder S Bhogal. 1997. The persona effect: affective impact of animated pedagogical agents. In *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*. 359–366.

[39] Shenzhen Wumii Technology Limited. 2020. Little English. https://apps.apple.com/cn/app/id1392803612. Accessed: 2020-10-05.

[40] Shanghai Liulishuo Information Technology Ltd. 2020. Liulishuo - Your Personal AI English Teacher. https://www.liulishuo.com/en/. Accessed: 2020-10-05.

[41] Wenting Ma, Olusola O. Adesope, J. Nesbit, and Q. Liu. 2014. Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology* 106 (2014), 901–918.

[42] Student Marketing. 2019. *Raising the profile in China of Australia's excellence in the delivery of English language training*. Technical Report. English Australia. https://www.englishaustralia.com.au/documents/item/585

[43] Nara M Martirosyan, Eunjin Hwang, and Reubenson Wanjohi. 2015. Impact of English proficiency on academic performance of international students. *Journal of International Students* 5, 1 (2015), 60–71.

[44] Hazel Morton, Nancie Gunson, and Mervyn Jack. 2011. Attitudes to subtitle duration and the effect on user responses in speech interactive foreign language learning. *Journal of Multimedia* 6, 5 (2011), 436–446. https://doi.org/10.4304/jmm.6.5.436-446

[45] Hazel Morton, Nancie Gunson, and Mervyn Jack. 2012. Interactive language learning through speech-enabled virtual scenarios. *Advances in Human-Computer Interaction* 2012 (2012). https://doi.org/10.1155/2012/389523

[46] Hazel Morton and Mervyn Jack. 2010. Speech interactive computer-assisted language learning: A cross-cultural evaluation. *Computer Assisted Language Learning* 23, 4 (2010), 295–319. https://doi.org/10.1080/09588221.2010.493524

[47] Noriko Nagata. 1993. Intelligent computer feedback for second language instruction. *The Modern Language Journal* 77, 3 (1993), 330–339.

[48] Noriko Nagata. 1995. An effective application of natural language processing in second language instruction. *CALICO Journal* 13, 1 (1995), 47–67.

[49] Noriko Nagata. 1996. Computer vs. workbook instruction in second language acquisition. *CALICO Journal* 14, 1 (1996), 53–75.

[50] Kae Nakaya and Masao Murota. 2013. Development and evaluation of an interactive English conversation learning system with a mobile device using topics based on the life of the learner. *Research and Practice in Technology Enhanced Learning* 8, 1 (2013), 65–89.

[51] Seyyed Ali Ostovar Namaghi, Seyyed Esmaail Safaee, and Abdolreza Sobhanifar. 2015. The effect of shyness on English speaking scores of Iranian EFL learners. *Journal of Literature, Language and Linguistics* 12 (2015), 22–28.

[52] Neasa Ní Chiaráin and Ailbhe Ní Chasaide. 2016. Chatbot technology with synthetic voices in the acquisition of an endangered language: Motivation, development and evaluation of a platform for Irish. In *LREC'16: Proceedings of the tenth international conference on language resources and evaluation*. 3429–3435.

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

[54] Test of English as a Foreign Language (TOEFL). 2020. The TOEFL Family of Assessments. https://www.ets.org/toefl. Accessed: 2020-10-05.

[55] Lourdes Ortega and Gina Iberri-Shea. 2005. Longitudinal Research In Second Language Acquisition: Recent Trends And Future Directions. *Annual Review of Applied Linguistics* 25 (2005), 26–45. https://doi.org/10.1017/s0267190505000024

[56] Sherry Ruan, Jiayu He, Rui Ying, Jonathan Burkle, Dunia Hakim, Anna Wang, Yufeng Yin, Lily Zhou, Qianyao Xu, Abdallah AbuHashem, Griffin Dietz, Elizabeth L. Murnane, Emma Brunskill, and James A. Landay. 2020. Supporting Children's Math Learning with Feedback-Augmented Narrative Technology. In *Proceedings of the Interaction Design and Children Conference* (London, United Kingdom) *(IDC '20)*. Association for Computing Machinery, New York, NY, USA, 567–580. https://doi.org/10.1145/3392063.3394400

[57] Sherry Ruan, Liwei Jiang, Justin Xu, Bryce Joe-Kun Tham, Zhengneng Qiu, Yeshuang Zhu, Elizabeth L. Murnane, Emma Brunskill, and James A. Landay. 2019. QuizBot: A Dialogue-Based Adaptive Learning System for Factual Knowledge. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13.

[58] Sherry Ruan, Angelica Willis, Qianyao Xu, Glenn M. Davis, Liwei Jiang, Emma Brunskill, and James A. Landay. 2019. BookBuddy: Turning Digital Materials Into Interactive Foreign Language Lessons Through a Voice Chatbot. In *Proceedings of the Sixth (2019) ACM Conference on Learning @ Scale* (Chicago, IL, USA) *(L@S '19)*. Association for Computing Machinery, New York, NY, USA, Article 30, 4 pages.

[59] Derek Sleeman and John Seely Brown. 1982. *Intelligent Tutoring Systems*. London : Academic Press. 345 pages pages.

[60] Tetyana Sydorenko, Tom F.H. Smits, Keelan Evanini, and Vikram Ramanarayanan. 2019. Simulated speaking environments for language learning: Insights from three cases. *Computer Assisted Language Learning* 32, 1-2 (2019), 17–48.

[61] International English Language Testing System. 2020. IELTS Home of the IELTS English Language Test. https://www.ielts.org/en-us. Accessed: 2020-10-05.

[62] Stergios Tegos, Stavros Demetriadis, and Thrasyvoulos Tsiatsos. 2014. A configurable conversational agent to trigger students' productive dialogue: A pilot study in the CALL domain. *International Journal of Artificial Intelligence in Education* 24 (2014), 62–91. https://doi.org/10.1007/s40593-013-0007-3

[63] Kurt VanLehn. 2011. The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems. *Educational Psychologist* 46, 4 (2011), 197–221. https://doi.org/10.1080/00461520.2011.611369

[64] Peter Vlugter, Alistair Knott, Jenny McDonald, and Craig Hall. 2009. Dialogue-based CALL: A case study on teaching pronouns. *Computer Assisted Language Learning* 22, 2 (2009), 115–131. https://doi.org/10.1080/09588220902778260

[65] Haining Wang, Russell Smyth, and Zhiming Cheng. 2017. The economic returns to proficiency in English in China. *China Economic Review* 43 (2017), 91–104.

[66] I Wang, Janet N Ahn, Hyojin J Kim, and Xiaodong Lin-Siegler. 2017. Why do international students avoid communicating with Americans. *Journal of International Students* 7, 3 (2017), 555–582.

[67] Lindy Woodrow. 2006. Anxiety and Speaking English as a Second Language. *RELC Journal* 37, 3 (2006), 308–328. https://doi.org/10.1177/0033688206071315

[68] Zhan.com. 2020. TOEFL Delta Fine Listening. http://top.zhan.com/toefl/listen/jingtingdelta.html. Accessed: 2020-10-05.

[69] Jiayi Zhu. 2020. Lanren English. https://apps.apple.com/cn/app/id730904222. Accessed: 2020-10-05.