Moving from off-the-shelf chatbots to a user-designed bespoke L2 chatbot

Dongkwang Shin, Gwangju National University of Education
Jang Ho Lee, Chung-Ang University

Abstract

This study investigates the application of Large Language Model (LLM)-based chatbots for second language (L2) learning, focusing on the three chatbot-building platforms such as ChatGPT, Poe AI, and Pi. Engaging 96 pre-service teachers in South Korea, it examined their perceptions of chatbots built via these platforms concerning human-likeness, pedagogical usefulness, and specific strengths and weaknesses. Participants were asked to create task-oriented chatbots using these platforms and to converse with them. The findings reveal varied perceptions of human-likeness among the chatbots, with Pi rated the highest. Regarding usefulness for L2 learning, the chatbots built via all three platforms were deemed beneficial, especially for engaging in realistic scenarios and providing authentic, context-appropriate expressions. Each platform demonstrated unique strengths but also showed some limitations, based on which we provide the pedagogical implications. Overall, the present study contributes to the evolving field of chatbot-assisted language learning, demonstrating the utility of LLM-based platforms in creating customized L2 learning chatbots.

Keywords: Artificial Intelligence, Chatbot, ChatGPT, Poe, Pi

Language(s) Learned in This Study: English


Introduction

Chatbot-assisted language learning has garnered much attention from the second language (L2) research community, as attested by the number of relevant review articles (e.g., Fryer et al., 2020; Huang et al., 2022; Lee et al., 2020) and meta-analyses (e.g., Bibauw et al., 2022; Lee & Hwang, 2022; Zhang et al., 2023). Research to date has revealed its social and pedagogical affordances for language learning, such as enhancing interpersonal communication, providing language knowledge and skill, to list just a few (see Huang et al., 2022 for a review of the extant primary studies on chatbot-assisted language learning). More recently, research has shifted its focus from chatbots built for general purposes to L2-focused researcher- and teacher-developed chatbots to maximize chatbot-assisted language learning efficacy (Çakmak, 2022; Zhai & Wibowo, 2022).

The recent development of Large Language Model (LLM) technology in the Artificial Intelligence (AI) domain has brought about significant changes to chatbot research, particularly regarding chatbots’ ability to comprehend human languages in context (e.g., Dan et al., 2023; de Winter, 2023). Meanwhile, the emergence of chatbot-building platforms, including ChatGPT (https://chat.openai.com/), Poe AI (https://poe.com/), and Pi (https://pi.ai/), which are all based on the aforementioned LLM technology, has made chatbot-building possible without programming knowledge. Thus, it is now possible for language teachers and learners to build their own chatbots, resulting in a paradigm shift in the field of chatbot-assisted language learning. However, there has been little research on these different platforms for building LLM-based chatbots, leaving their pedagogical values unexplored.
To this end, in the present study, 96 pre-service teachers were asked to engage in chatbot-building activities via the three platforms (i.e., ChatGPT, Poe AI, Pi), and evaluate the chatbots generated with these platforms concerning their human-likeness, pedagogical usefulness, and strengths and weaknesses. This study aimed to contribute to the field of chatbot-assisted language learning by gleaning pre-service teachers’ perception of the newly emerging LLM-based chatbot-building platforms and by explicating the pedagogical implications for L2 teaching communities.

Background

In this section, we first review earlier research on chatbots employed in L2 learning. Next, we introduce more recent studies on L2 chatbots, which are developed based on chatbot-building platforms and LLM. This section concludes by addressing the research questions for the current study.

Earlier L2 Research on Chatbots

The idea of using computer programs as conversational practice partners emerged more than two and a half decades ago (Atwell, 1999), with a thorough examination of the potential of chatbots for L2 instruction having begun with Coniam’s studies on the chatbots available at the time (Coniam, 2004; Coniam, 2008; Coniam, 2014). Most of the chatbots included in Coniam’s studies (e.g., ALICE AI Foundation’s Dave, Lucy, Elbot) were award-winners at the Chatterbox Challenge and the Loebner Prize, and their ability to mimic human conversational patterns made them a focus of the research, given their potential as L2 practice partners. Notably, most chatbots (e.g., Eliza, Jabberwacky) addressed in early studies, including Coniam’s, were not specifically designed for language learning; rather, they were chatbots built for non-educational purposes.

In the 2010s, L2 researchers employed two different types of chatbot, the first being another award-winning chatbot (e.g., Cleverbot, Mitsuku) built for general purposes (e.g., Kim, 2018; Fryer et al., 2019; Gallacher et al., 2018) and those developed based on a self-designed system (e.g., Goda et al., 2014; Jia et al., 2012). Overall, these studies revealed chatbots’ strengths (e.g., providing opportunities to practice L2, linguistic assistance, and level-appropriate input) as well as weaknesses (e.g., unable to comprehend L2 learners’ input, providing irrelevant utterances occasionally) (see Huang et al., 2022 for a recent review). Despite the various strengths identified, it is noteworthy that most chatbots introduced in this era of L2 research may still have limited pedagogical efficacy, as L2 teachers may be unable to participate in chatbot building. More recent L2 research has sought to overcome this limitation by considering accessible chatbot building platforms, which we turn to next.

Platform- and LLM-Based Chatbots for L2 Learning

From the early 2020s on, researchers began employing user-friendly platforms to develop their own chatbots, with Google’s Dialogflow (https://cloud.google.com/dialogflow) being the most extensively used (e.g., Kim et al., 2022; Kohnke, 2023; Lee & Jeon, 2022) for this purpose. A major strength of chatbots developed via such platforms is that they could be designed to perform a specific L2 task (e.g., making an appointment with foreign friends, ordering a beverage at a coffee shop), with which learners could engage in purposeful communication. Also, it is notable that these platforms provide accessible chatbot-building environments for language teachers who have no programming knowledge, who could design their own chatbots given their students’ needs and proficiency levels. When developing chatbots based on these platforms, designers are expected to determine a set of linguistic expressions (or functions) for learners to practice, and then contextualize such expressions into the platform given a pre-determined context (e.g., ordering at a coffee shop) by entering a set of expected user utterances and chatbot responses (see Kim et al., 2022 for details). Although such chatbots may be more pedagogically applicable than general-purpose ones (e.g., Cleverbot), they may have some constraints, namely that, when learners make utterances irrelevant to the predetermined contexts and use unexpected expressions not entered into the platform, chatbots may fail to respond appropriately.
The recent development of Chat Generative Pre-trained Transformer (ChatGPT, henceforth) based on an LLM by OpenAI (2022) has opened up new avenues for chatbot research in various domains (Ghumra, 2022), including L2 research (e.g., Kohnke et al., 2023; Lee et al., 2023). The most noticeable advantage of LLM-based chatbots is that they can overcome the aforementioned problems. Various uses of LLM-based chatbots have recently been introduced to the L2 literature, highlighting their ability to respond to L2 reading comprehension questions (e.g., Ahn, 2023; Kwon & Lee, 2023), generate L2 testing items (e.g., Shin, 2023; Shin & Lee, 2023), and to produce content that is often indistinguishable from human-produced language (e.g., Casal & Kessler, 2023; Guo et al., 2023), to list a few. As a chatting partner, LLM-based chatbots have shown much enhanced capability to understand users’ utterances in contexts (Lee et al., 2023), and thus may better prevent communication breakdown. Moreover, from the developer’s point of view, creating LLM-based chatbots can be much easier, as it can be done simply by entering prompts into the platform system, which is now available in ChatGPT and Poe AI. This feature largely contrasts from developing chatbots via platforms like Dialogflow, in which developers need to type several sets of (expected) utterances.

**Research Questions**

Despite this growing interest in applying LLM-based chatbots in L2 instruction, little effort has been made to develop LLM-based chatbots for specific tasks for L2 practice. Given such a significant gap in the field, the present study aimed to examine the potential of three different chatbot-building platforms (i.e., ChatGPT 3.5, Poe AI, Pi) for developing LLM-based chatbots for L2 learning, especially when they are designed and developed by users (i.e., pre-service teachers).

To this end, the present study addresses the following three research questions:

Research Question 1: What are pre-service teachers’ perceptions of the LLM-based chatbots developed via three different platforms regarding their human-like qualities?

Research Question 2: What are pre-service teachers’ perceptions of the LLM-based chatbots developed via three different platforms regarding their usefulness for English learning?

Research Question 3: What are the relative strengths and weaknesses of chatbots developed via each platform?

**Method**

**Participants**

This study involved 96 pre-service teachers, 30 males and 66 females, enrolled in a course on computer-assisted language learning at a university of education in South Korea. The pre-service teachers were all undergraduate students between the ages of 20 and 24, who were being trained to become elementary school teachers. These participants should have more in-depth knowledge about educational theories and technology-enhanced language learning than ordinary EFL students, given their institution’s curriculum and their educational backgrounds. The participants were given an introduction to the purpose of the study prior to participation, and were asked to withdraw from the study if they wished. The participants who completed the target task received extra credit for the course they attended.

**Chatbot-Building Platforms**

As mentioned above, this study utilized three different platforms (i.e., ChatGPT 3.5, Poe AI, Pi) for developing chatbots for L2 learning. The main reason for selecting these platforms was that they are all free to use, and each platform has its own unique features for building customized chatbots. Instructions on how to use each platform were provided in a worksheet with screen captures and a chatbot creation manual with instructor explanations. One of the researchers created a chatbot based on the example in the manual and demonstrated it to the participants, who then followed the demonstration and piloted the same chatbot. After 30 minutes of instruction and practice, the participants were asked to develop a chatbot on each platform.
with their own ideas and communicate with the customized chatbots for two hours. All tasks were completed in class.

The first platform was ChatGPT 3.5 (OpenAI, 2022). Although the paid version of ChatGPT 4.0 offers a chatbot development feature called My GPTs, making it easy to develop a custom chatbot, we used ChatGPT 3.5 because this study aimed to use only free chatbot development platforms, in consideration of accessibility for general users. In doing so, we utilized a feature called ‘Custom instructions’, which is also available in the free version of ChatGPT 3.5. Figure 1 illustrates how one can enter prompts regarding the target conversation context and how to respond in the conversation in ‘Custom instructions’ so that ChatGPT can act as a conversational partner for L2 learners. As one of the tasks given to the pre-service teachers, they created their own task-oriented chatbot by entering the desired conversation situation and utterance conditions. Specifically, when using ‘Custom Instructions,’ pre-service teachers entered missions for the chatbot in the top box, including the chatbot’s role and the conversational context. In the bottom box, participants asked the chatbot to respond to them in two sentences or fewer, as ChatGPT’s responses are typically too long for informational purposes; they also asked the chatbot to express itself at a specific grade level (e.g., K1–12).

**Figure 1**

*An Example of Prompts Entered in ‘Custom Instructions’ in ChatGPT*

The second platform, Poe AI (Quora, 2022), offers chatbot services based on various free and paid LLMs, including ChatGPT 3.5, ChatGPT 4.0, Claude, Llama, PaLM, and Gemini-Pro. Poe AI launched the ‘Create bot’ feature before releasing the aforementioned feature called My GPTs in ChatGPT 4.0. The process of creating a chatbot on the Poe AI platform is similar to using the chatbot creation template on My GPTs, as illustrated below in Figure 2.
Figure 2

The Process of Creating a Customized Chatbot on the Poe AI Platform

**handle**  
*Must contain 4-20 characters, including letters, numbers combinations.*  
AirlineTicketingBot

**Basic bot**  
ChatGPT

**prompt**  
*Teach your bot how to behave and respond to user messages.*  
See best practices for prompts?  
You are a chatbot for learning English and you play the role of a ticket agent for Kor Airline. You can suggest different priced tickets for different dates, and you can also advise on seat assignments. You can also create a scenario to explain Kor Airlines’ baggage weight policy. However, please do not make your answer too long (e.g. within 2 sentences) and use an English level of K-10 (17 years old) or below.

**knowledge base**  
*Provide custom knowledge for the bot to reference in providing responses.*  
knowledge base for relevant sections based on user messages. Data from made visible to other users through bot responses or citations.  
+ Add knowledge sources

**greeting message**  
The bot sends this message at the beginning of every conversation.  
Hello, welcome to Kor Airline! What can I help you?

**advanced**  

- Answer suggestions
- Rendering Markdown content
- custom temperature  
*Control the creativity of your bot’s responses.*  
Default: 0.35

---

**Step 1:** Enter the name of your chatbot without spaces.

**Step 2:** Choose an LLM to power your chatbot.

**Step 3:** Enter all the directions and requirements, including the chatbot’s role, the conversational context, and how the chatbot should respond.

**Step 4:** Upload a document as a PDF file with examples of specific conversations or with information too large to type as prompts, using some sort of machine learning if necessary.

**Step 5:** Enter the chatbot’s opening statement at the beginning of a conversation, if needed.

**Step 6:** Select ‘Answer suggestions’ when recommending conversational expressions to learners, and control how creative the chatbot should be in ‘Custom temperature (the higher, the more creative).’
bot profile

About Me
This chatbot is an English task-based chatbot for a fictional airline that is designed to learn expressions related to booking and purchasing flight tickets.

access

Make your bot publicly accessible
When this setting is enabled, the bot will be added to your profile and will be publicly accessible. Turning this off will make your bot private.

monetization

Visit Creators in the menu to learn more about Creator Monetization and unlock exclusive tools.

Step 7: Introduce your chatbot to users, including what it is for (optional).

Step 8: Decide if it’s available for the public to discover and use.

Step 9: Click ‘Save’ to finish developing your chatbot.

Notably, although Poe AI’s chatbot may be less powerful than the one developed via My GPTs (GPT 4.0), Poe AI’s represents a significant technological advancement in enabling natural conversation on unlimited topics, compared to existing scenario-based chatbots (i.e., developed via the platform like Dialogflow).

The third and last platform was Pi (InFlaction AI, 2023), which is known as one of the top three LLM-based chatbots in the world regarding performance, along with ChatGPT and Claude. Although Pi does not have a feature of building chatbots like the other two, it has features like ‘venting,’ ‘brainstorming ideas,’ ‘relationship advice,’ ‘discovering books, movies, or podcasts,’ ‘practicing for a big conversation,’ and discussing a wide range of topics. With Pi, participants were asked to employ the ‘Practice a big conversion’ feature. They were then instructed to type the chatbot’s role and situation for the role-playing, just the same way as they had done in the previous two tasks with ChatGPT 3.5 and Poe AI. Pi occasionally refused to role-play, but when participants explained that they wanted to learn a foreign language, Pi then willingly participated. Figure 3 illustrates how to set up roles and situations for role-playing on this platform.

Instrument

After developing the chatbots via the three platforms and engaging conversationally with their own chatbots, the pre-service teachers were given a post-task survey through a Google Forms, which consisted of the Likert-scale items on each chatbot’s human-like qualities and usefulness for L2 learning, as well as follow-up questions for these items. Also, the survey included an open-ended question about each chatbot’s relative strengths and weaknesses developed via different platforms. Afterward, participants were instructed to copy and paste their conversation logs with each chatbot into a separate Word file and submit it to the first author, ensuring they developed the customized chatbots and engaged conversationally as instructed. The structure of the post-task survey and relevant research questions are provided in the Appendix A.

Data Analysis

The questionnaire items included in the survey were analyzed both quantitatively and qualitatively. The Likert-scale and multiple-choice items in the survey (i.e., #1, #2, #3, #5, #6, #7, #9, #10, #11) were analyzed descriptively. That is, the participants’ responses to each category in each item were calculated, and summarized. Additionally, repeated measures ANOVA, preceded by sphericity assumption test (i.e., Mauchly’s test), was used to analyze the Likert-scale items on LLM chatbots’ human-like qualities and usefulness for L2 learning.

In the case of the open-ended items (i.e., #4, #8, #12, #13), the responses that were not specifically relevant to the questionnaire item (e.g., “I really liked chatbot-building tasks”) or were too short or general to capture the intended meaning (e.g., “they were very useful for English learning”) were first excluded. The
remaining responses were then read by the authors several times until they could be grouped into a list of themes (e.g., authenticity of dialogue, context-appropriate expressions, built-in features, formality of chatbot language tone).

**Figure 3**

*Setting Up Roles and Situations for Role-Playing on the Pi Platform*

---

**Results**

**Perceptions of the LLM-Based Chatbots in Terms of Their Human-Like Qualities**

Our first survey item asked about the pre-service teachers’ perceptions of the LLM-based chatbots developed via these three platforms in terms of their human-like qualities when engaged conversationally. As Table 1 indicates, participants’ average rating of the three chatbots varied, ranging from 3.68 for ChatGPT to 4.08 for Pi.

The results of the Mauchly’s test revealed that the assumption of sphericity was violated, \( x^2(2) = 9.76, p = .008 \). Given this, Greenhouse-Geisser corrected value was interpreted. It was found that the participants’ perception of human-like qualities of chatbot was significantly affected by the platform, \( F(1.82, 172.95) = 4.56, p = .014 \), partial \( \eta^2 = 0.05 \). The results of post hoc tests (Tukey) further revealed that Pi was rated significantly higher than ChatGPT 3.5 (\( t = 2.74, df = 95, p < .05 \)), but there was no significant difference between Pi and Poe AI (\( t = 1.69, df = 95, p > .05 \)), and between ChatGPT 3.5 and Poe AI (\( t = 1.48, df = 95, p > .05 \)). Still, it was found that over 70% of participants opted for either ‘5 = indistinguishable from humans’ or ‘4 = very similar to humans’ for this Likert-scale item for Pi. Participants gave their reasons in the order of ‘authenticity of the content of the dialogue’ (\( n = 51 \)), ‘chatbot’s capacity to comprehend the dialogue in contexts’ (\( n = 48 \)), and ‘chatbot’s natural expressions’ (\( n = 46 \)).
Table 1

Participants’ Perceptions of the Three Chatbots Regarding Their Human-Like Qualities (n = 96)

<table>
<thead>
<tr>
<th></th>
<th>1 (1.0%)</th>
<th>2 (20.8%)</th>
<th>3 (6.3%)</th>
<th>4 (53.1%)</th>
<th>5 (18.6%)</th>
<th>M(SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChatGPT 3.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.68 (1.04)</td>
</tr>
<tr>
<td>Poe AI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.84 (0.92)</td>
</tr>
<tr>
<td>Pi</td>
<td>2 (2.1%)</td>
<td>8 (8.3%)</td>
<td>11 (11.5%)</td>
<td>34 (35.4%)</td>
<td>41 (42.7%)</td>
<td>4.08 (1.03)</td>
</tr>
</tbody>
</table>

Notes. Likert scale (5 = indistinguishable from humans, 4 = very similar to humans, 3 = moderate, 2 = communication is possible, although it occasionally produces some awkward expressions, 1 = difficult to manage a conversation in context)

Perceptions of the LLM-Based Chatbots Regarding Their Usefulness for L2 Learning

The next questionnaire item concerned participants’ perceptions of the three chatbots regarding their usefulness for L2 learning. In Table 2, participants’ average rating of the three chatbots were similar, ranging from 3.89 for Poe AI to 3.97 for ChatGPT 3.5 and Pi.

The results of the Mauchly’s test revealed that the assumption of sphericity was violated, $x^2(2) = 14.58, p < .001$; therefore, Greenhouse-Geisser corrected value was interpreted. It was found that the participants’ perception of LLM-based chatbots’ usefulness for L2 learning was not significantly affected by the platform, $F(1.75, 166.14) = 0.45, p = .62$, partial $\eta^2 = 0.005$. Because the main effect (i.e., platform) was not significant, we did not perform any further post hoc tests. Overall, it was found that over 70% of participants judged all three chatbots to be useful for L2 learning.

Table 2

Participants’ Perceptions of the Three Chatbots Regarding Their Usefulness for L2 Learning (n = 96)

<table>
<thead>
<tr>
<th></th>
<th>1 (2.1%)</th>
<th>2 (4.2%)</th>
<th>3 (16.7%)</th>
<th>4 (49.0%)</th>
<th>5 (28.1%)</th>
<th>M(SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChatGPT 3.5</td>
<td></td>
<td>4 (4.2%)</td>
<td>18 (18.8%)</td>
<td>51 (53.1%)</td>
<td>23 (24.0%)</td>
<td>3.97 (0.77)</td>
</tr>
<tr>
<td>Poe AI</td>
<td></td>
<td>5 (5.2%)</td>
<td>20 (20.8%)</td>
<td>52 (54.2%)</td>
<td>19 (19.8%)</td>
<td>3.89 (0.78)</td>
</tr>
<tr>
<td>Pi</td>
<td></td>
<td>2 (2.1%)</td>
<td>4 (4.2%)</td>
<td>16 (16.7%)</td>
<td>47 (49.0%)</td>
<td>3.97 (0.90)</td>
</tr>
</tbody>
</table>

Notes. Likert scale (5 = very useful, 4 = useful, 3 = moderate, 2 = not useful, 1 = does not have much pedagogical value)

Regarding the examined chatbots’ usefulness, most participants pointed to the feature that students could engage in the scenario (context) they designed, which they found to be very realistic, as the following excerpts showed:

The [chatbot’s] ability to apply different scenarios is very appealing. Previously, I had to learn by looking at examples in textbooks and reference books, but if I can write my own scenarios, I think it would be very helpful to be able to use English in various contexts that I can use in real life. (Participant #13)
This is because it mimics real-life situations very closely. For example, when I presented an airline ticketing situation and practiced having a conversation with the chatbot, I found it helpful because the level of interaction was very similar to real life. (Participant #32)

It was interesting to assign a role to the chatbot; it could find the answer [to a question] by itself and could answer my questions accordingly. I think it would be fun to have a chatbot like this in English classes in the future. (Participant #95)

Another frequently-mentioned advantage of these chatbots for L2 learning concerned their authentic and context-appropriate expressions, which are sometimes not available in textbooks and teachers’ input: “They [Chatbots] use words that can be used in real life situations, so it feels like a real conversation (Participant #39),” “I think their expressions are very authentic (Participant #10),” and “They [Chatbots] give a lot of useful and natural English expressions (Participant #90).”

Overall, participants reported that these chatbots’ value was that they could provide appropriate contextualized tasks for L2 learning, in which participants could engage in purposeful communication and receive authentic L2 input.

Relative Strengths and Weakness of Each LLM-Based Chatbot

This section will address participants’ perceived strengths and weaknesses of each LLM-based chatbot. First, ChatGPT 3.5’s major strengths, as perceived by the pre-service teachers, were its capacity to provide a large amount of data and useful answers. Also, it was found to offer a wide range of pedagogically useful extensions (e.g., looking up words, providing Internet resources and language practices), as seen in the following excerpts: “ChatGPT is one of the most well-known platforms and has many extensions that can be used to teach English (Participant #64),” and “ChatGPT seems to be very good at answering users’ questions based on a lot of data and at providing educationally useful answers (Participant #31).”

However, ChatGPT was also found to have a formal tone when acting as a conversational partner, and the amount of input it provided could sometimes be considered overwhelming: “ChatGPT still feels like a manual (Participant #56),” “ChatGPT’s answers feel stiff and dictionary-like sometimes. It needs to be improved so that its answers mirror those used in real conversations (Participant #80),” and “It offers too much information, so it was a bit difficult to interpret (Participant #93).”

Poe AI was perceived positively concerning its built-in features, such as ‘custom temperature’ (i.e., controls the creativity of the chatbot’s utterances) and ‘answer suggestions’ (i.e., offers a range of responses for users to choose), and its accessible user-interface, with which language teachers or learners could build their own chatbots easily: “I think Poe AI is best suited for building purposeful chatbots because it gives you a lot of control over how creative you can be, which is very helpful as an educational feature for students (Participant #74),” “I love that Poe AI has an intuitive user-interface that makes customizing a chatbot quick and easy (Participant #63),” and “Poe AI is very easy to use because it’s a platform set up for situational dialogues, and the ‘Answer suggestions’ feature is very helpful for English learners (Participant #66).”

Meanwhile, Poe’s answers were considered verbose by several participants, which may limit its use for learners with lower levels of English proficiency: “I don’t think I’d actually use this [POE’s] sentence in conversations. It feels like it’s stretching a sentence that could be said concisely (Participant #52),” and “The text is too long, making it tiring to read its utterances sometimes, compared to other chatbots (Participant #53).”

Finally, Pi was considered the most accessible for building chatbots, as it does not require users to type prompts (e.g., Pi was easiest to use because it didn’t require any setup, cf. Participant #56). Additionally, several participants reported positively on Pi’s human-like qualities in managing English conversation with users, which accords with the finding related to the Likert-scale item presented above. The relevant responses from participants were as follows: “I thought I was talking to a real person, not only because of the friendly responses, but also because of the emoticon use. The level of understanding and empathy was very impressive (Participant #14).”
It really felt like I was talking to a real native speaker, especially the emoticons used and the English expressions . . . I felt like I was talking to a human being and not a robot, which made the whole conversation exciting and enjoyable. (Participant #21)

Pi, however, was also perceived to have some limitations, as it gives little control to users regarding the target context, language levels, and roles assigned to the chatbots, as the following excerpts illustrate: “I think Pi lacks details compared to ChatGPT and Poe in the setup part (Participant #93),” “In the case of Pi, I was disappointed because I could not adjust chatbot’s language expression levels (Participant #62),” and “I felt uncomfortable with the fact that there was no way for users to set up the chatbot to get the desired conversation (Participant #65).”

Reflection and Implications

This study examined pre-service teachers’ perceptions of three chatbots (ChatGPT, Poe AI, Pi) developed via different platforms for L2 learning. Our findings align with recent L2 research, which has suggested the strong potential of LLM-based chatbots for L2 learning and instruction (e.g., Kohnke et al., 2023; Kwon & Lee, 2023; Lee et al., 2023; Shin & Lee, 2023). In particular, it accords with the recent finding that LLM-based chatbots are superior in quality in terms of their human-like qualities (Lee et al., 2023), thus overcoming the limitations of chatbots based on earlier technologies (e.g., unable to comprehend users’ input) (e.g., Gallacher et al., 2018). Overall, the emergence of accessible chatbot-building platforms and LLM technology development are expected to significantly advance chatbot-assisted language learning, as demonstrated by the present study’s results.

It was further found that each chatbot has its own strengths and weaknesses. Namely, ChatGPT provides informative responses for learning, but participants found the tone to be very formal. Poe AI can be customized to meet different needs and can suggest answers, but can be verbose according to the open-ended responses of the sampled pre-service teachers. Finally, Pi is perceived by participants as the most human-like, but its settings are not very customizable.

Based on the findings regarding the sampled pre-service teachers’ perceptions of the three chatbots, we offer the following suggestions on how each chatbot could be used differently for L2 teachers. First, given that Pi does not require user prompting and provides the most authentic conversational experiences with L2 speakers, it can be used for the purpose of developing learners’ fluency in a variety of everyday communicative situations without specific target expressions in mind. Poe AI could be used when teachers want to design and implement a specific L2 task with a set of target expressions by entering a detailed prompt when building a target conversational agent. It is noteworthy that Poe AI’s language could be modified in terms of creativity (by adjusting the ‘custom temperature’) and vocabulary and sentence structure (via prompts), and thus language teachers could even build customized chatbots for different levels of students within their classes. Poe AI could be particularly useful for learners with relatively low levels of L2 proficiency and willingness to communicate, thanks to its built-in ‘Answer suggestions’ feature. ChatGPT, while similar to Poe AI (i.e., building chatbots via prompts), could be used to build a supplementary interaction-based tool for skill-building tasks, given its formal tone and informative responses. For example, in reading tasks, learners could use ChatGPT’s conversational agent to ask any question that requires a high level of inference (e.g., to understand the author’s intent or to figure out the underlying meaning of the idiomatic phrases in the passage), or in post-reading tasks that require learners to find more information about the target passage (e.g., find information about other types of pollution not discussed in the passage, and find any similarities and differences between different types of pollution).

The present study is not without limitations. First, it was an observational study of participants’ perceptions rather than an experimental study of the effects of LLM-based chatbots. It remains unknown whether learners could actually benefit from these LLM-based chatbots considering the current technological developments. Therefore, we call for an experimental study that aims to investigate the effectiveness of
using LLM-based chatbots, ideally comparing different chatbot types. Second, this study’s participants were pre-service teachers, who were selected because of their knowledge of educational theories and technology-enhanced language learning. Therefore, the reported views in the present study may differ from those of students with limited levels of L2 proficiency.

Despite the aforementioned limitations, we believe that our introduction to three different chatbot-building platforms and findings related to each platform’s strengths and weaknesses are expected to be highly useful for L2 teaching practitioners. Also, this study contributes to the field of chatbot-assisted language learning by demonstrating the potential of these three platforms for building customized LLM-based L2 chatbots, which could catalyze the next phase of chatbot-assisted language learning research.

Acknowledgements

We thank the section editor and reviewers for their insightful feedback on our manuscript. We also thank the students who participated in our study.

References


Inflection AI. (2023). *Pi [Large Language Model]*. https://pi.ai/


Appendix A. The Structure of the Post-Task Survey

<table>
<thead>
<tr>
<th>Item number</th>
<th>Platform</th>
<th>Item type</th>
<th>Item content</th>
<th>Relevant RQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ChatGPT 3.5</td>
<td>Likert (5 scale)</td>
<td>The degree of chatbot’s human-likeness</td>
<td>RQ1</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>Multiple choice</td>
<td>Rationale for their choice for item #1</td>
<td>RQ1</td>
</tr>
<tr>
<td>3</td>
<td>Poe AI</td>
<td>Likert (5 scale)</td>
<td>The degree of chatbot’s usefulness of L2 learning</td>
<td>RQ2</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>Open-ended</td>
<td>Rationale for their choice for item #3</td>
<td>RQ2, 3</td>
</tr>
<tr>
<td>5</td>
<td>Poe AI</td>
<td>Likert (5 scale)</td>
<td>The degree of chatbot’s human-likeness</td>
<td>RQ1</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>Multiple choice</td>
<td>Rationale for their choice for item #5</td>
<td>RQ1</td>
</tr>
<tr>
<td>7</td>
<td>Poe AI</td>
<td>Likert (5 scale)</td>
<td>The degree of chatbot’s usefulness of L2 learning</td>
<td>RQ2</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>Open-ended</td>
<td>Rationale for their choice for item #7</td>
<td>RQ2, 3</td>
</tr>
<tr>
<td>9</td>
<td>Poe AI</td>
<td>Likert (5 scale)</td>
<td>The degree of chatbot’s human-likeness</td>
<td>RQ1</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>Multiple choice</td>
<td>Rationale for their choice for item #9</td>
<td>RQ1</td>
</tr>
<tr>
<td>11</td>
<td>Poe AI</td>
<td>Likert (5 scale)</td>
<td>The degree of chatbot’s usefulness of L2 learning</td>
<td>RQ2</td>
</tr>
<tr>
<td>12</td>
<td></td>
<td>Open-ended</td>
<td>Rationale for their choice for item #11</td>
<td>RQ2, 3</td>
</tr>
<tr>
<td>13</td>
<td>All three chatbots</td>
<td>Open-ended</td>
<td>Strengths and weaknesses of each chatbot (in comparison with the other ones)</td>
<td>RQ3</td>
</tr>
</tbody>
</table>

About the Authors

Dongkwang Shin received his PhD in Applied Linguistics from Victoria University of Wellington and is currently a professor at Gwangju National University of Education, Republic of Korea. His research interests include corpus linguistics, CALL, and AI-based language learning.

E-mail: sdhera@gmail.com  
ORCiD: https://orcid.org/0000-0002-5583-0189

Jang Ho Lee received his DPhil in education from the University of Oxford, and is presently a Professor at Chung-Ang University, South Korea. His areas of interest are AI-based language learning, L1 use in L2 teaching, and vocabulary acquisition. Jang Ho Lee is the corresponding author. All correspondence regarding this publication should be addressed to him.

E-mail: jangholee@cau.ac.kr  
ORCiD: https://orcid.org/0000-0003-2767-3881