Language Learning & Technology
ISSN 1094-3501 CC BY-NC-ND

February 2025, Volume 29, Issue 2
pp. 132–160

ARTICLE

Using an AI-powered chatbot for improving L2 Korean grammar: A comparison between proficiency levels and task types

Ji-young Shin, University of Toronto Mississauga*

Yujeong Choi, University of Toronto

Abstract

The use of AI-powered chatbots has recently been extensively examined for second language (L2) learning. While their positive effects have been widely reported regarding L2 English learning, studies involving less commonly taught languages (LCTLs) are scant. The current study incorporated an AI chatbot called Iruda in L2 Korean teaching, to improve Korean lexico-grammar, while examining moderation of proficiency levels and task types. Sixty-six students engaged with AI chatbot-incorporated tasks with half of the grammar points of the course curriculum throughout the semester. During their final exam, students' grammar was assessed through selected-response (SR) and constructed-response (CR) tasks (sentence completion and sentence composition). Using a paired-sample t-test, we compared item mean differences between grammatical features practiced with chatbot-incorporated and conventional methods. Items related to chatbot practice showed a significantly higher mean ($t(65) = 4.29$, $p < .0001$). Follow-up regression and repeated-measures ANOVAs revealed greater changes by chatbot practice among students with lower pre-existing Korean proficiency and significant moderation of task types with higher SR item means. Students' perceptions were also positive. The study demonstrated the effectiveness of chatbot-enhanced instruction for LCTLs, while suggesting nuanced application for learners of varied proficiency levels and different task types.

Keywords: *AI chatbots*; *Less Commonly Taught Languages*; *Teaching Grammar*; *Korean as a Foreign Language*

Language(s) Learned in This Study: *Korean*

APA Citation: Shin. J., & Choi, Y. (2025). Using an AI-powered chatbot for improving L2 Korean grammar: A comparison between proficiency levels and task types. *Language Learning & Technology*, *29*(2), 132–160. https://hdl.handle.net/10125/73614

Introduction

With the recent introduction of generative artificial intelligence (AI) chatbots built on a large language model (LLM) such as ChatGPT (OpenAI, 2023), there has been a surge of studies exploring the efficacy of conversational AI for language teaching and learning (Bibauw et al., 2022; Fryer et al., 2020; Hwang & Chang, 2023; Jeon et al., 2023; Zhang et al., 2023). Teachers and students across diverse learning contexts have reported positive perspectives of using AI chatbots along with their effectiveness in improving second language (L2) proficiency, encompassing both productive and receptive skills, although little is known about grammar teaching (Chen et al., 2020; Lee & Jeon, 2024; Jeon et al., 2023; Tai & Chen, 2024; Zhang et al., 2024; Zhang & Zou, 2022). The positive pedagogical impact of AI chatbots is primarily due to their conversational functionality and accessibility. Conversations with virtual agent-generated interactions are fairly analogous to those in human language exchanges although there are some variations in pragmatic behaviors (Ockey & Chukharev-Hudilainen, 2021; Timpe-Laughlin et al., 2024). Therefore, language

*** Corresponding Author:** Ji-young Shin, jiyoung.shin@utoronto.ca

practice through tasks using conversational AI potentially serve as competent alternatives of human interlocutors, complementing traditional teaching. Considering that conversations facilitate L2 learning through socio-cognitive interaction opportunities with meaning negotiation (Long, 2015), human-like conversational functionality of AI chatbots can improve L2 learning. Along with conversational functions, AI chatbots have an advantage over human conversants regarding their accessibility because they are available anytime, anywhere, often free of charge, with one click (Chen et al., 2020; Lee & Jeon, 2024).

These characteristics of AI chatbots can be particularly beneficial for teaching and learning less commonly taught languages (LCTLs), considering their relatively fewer and less accessible interaction opportunities compared with commonly taught languages such as English, Spanish (in the US) and French (in Canada or Africa). We acknowledge that increased connectivity through technological developments and globalization has certainly helped narrow this gap, particularly through online platforms (e.g., YouTube, Zoom) (Isbell, 2018). Nevertheless, live, interactive language practice opportunities for LCTL learners remain less abundant than those available to learners of commonly taught languages, which may elevate the role of supplementary interaction with AI chatbots by offering consistent, on-demand practice. However, AI chatbot research has predominantly concerned L2 English (Jeon et al., 2023; Zhang et al., 2023), providing minimal insights into AI chatbots' potential for teaching LCTLs, such as L2 Korean. Additionally, pedagogical benefits of technology-mediated instruction are generally moderated by learner proficiency and task characteristics (Lin, 2014; Ziegler, 2016), but these moderators have rarely been examined in AI chatbot-enhanced L2 instruction.

The current study developed AI chatbot-incorporated tasks in an L2 Korean course and evaluated their impact on improving knowledge and use of L2 Korean grammar, which students reported as most challenging, while examining moderation of learner proficiency and grammar task types along with student perceptions. Given no published studies (to our best knowledge) on AI chatbot-enhanced L2 Korean teaching, along with the limited number of studies on technology-enhanced L2 Korean teaching and learning (Isbell, 2018; Nishioka, 2023; Sato et al., 2015), the study will make meaningful implications for technology-enhanced L2 pedagogy, specifically for the L2 Korean context.

Literature Review

Characteristics of Chatbot-Human Conversations

Chatbots typically refer to digital entities that engage with users through natural human language (Fryer et al., 2020). While acknowledging affordances can vary depending on specific features of various types of chatbots, it is their interactional and conversational capabilities that lead to positive pedagogical effects of chatbots on L2 learning (Bibauw et al., 2022; Fryer et al., 2020). Overall, chatbots or computerized dialogue systems appear to elicit language performance comparable to that observed during face-to-face (FTF) human-to-human conversations. In L2 English assessment settings, Timpe-Laughlin, Sydorenko, and Dombi (2022) and Ockey and Chukharev-Hudilainen (2021) compared linguistic performance in human-partner and computer-partner conditions, during role-play and paired oral discussion tasks, respectively. Both studies confirmed similar linguistic performance in grammar, vocabulary, fluency, and pronunciation between the two conditions. Additionally, Forsyth et al. (2019) found higher word count in L2 performance during chatbot-to-human conversations than human-to-human conversations. These findings support the use of chatbots as an alternative of human interlocutors for practicing L2 conversations.

Pragmatic behaviors involved in conversations with digital agents, however, showed slight differences. Compared with FTF human-to-human exchanges, conversations with digital partners demonstrated lower interactional competence (Ockey & Chukharev-Hudilainen, 2021) and lower social rapport as reflected in more direct requests (Timpe-Laughlin & Dombi, 2020) and fewer instances of turn-taking, clarification, and backchanneling (Timpe-Laughlin et al., 2024). Similar implications might apply to comparisons with synchronous computer-mediated communication (SCMC) among humans, given the minimal differences between FTF and SCMC contexts (Ziegler, 2016). These pragmatic differences underscore the importance

of designing socially engaging platforms and tasks to maximize the pedagogical effects of using AI chatbots.

The Use of AI-Powered Chatbots in L2 English Education

Teacher and Student Perceptions and Attitudes

Given the novelty of the technology behind AI chatbots, the most researched topic was teachers' and learners' perceptions and attitudes towards using AI chatbots for L2 teaching and learning (Jeon et al., 2023). Overall, teachers' perceptions of conversational AI were positive in both English as a second language (ESL) and foreign language (EFL) contexts. For example, EFL teachers in Korea recognized positive roles of ChatGPT as L2 conversant content resources, and teaching and evaluation assistant (Jeon & Lee, 2023). Teachers in a US intensive English program reported similar positive perceptions, specifically regarding the flipped classroom model (Timpe-Laughlin et al., 2022). Aligned with teachers' views, learners' perceptions and attitudes were also positive across different EFL learning contexts, learner ages, and chatbot types. Elementary school students using a voice-controlled conversational agent in Korea (Lee & Jeon, 2024) and college students interacting with Google Assistant in Taiwan (Chen et al., 2023) reported their positive learning experiences and perceived usefulness of the agents, mainly due to their human-like linguistic and social performance. However, some variations were also noted between different English proficiency levels—more proficient English users reported more successful chatbot-human communication (Chen et al., 2023). Learners' positive responses support the potential usefulness of AI chatbots across diverse L2 learning contexts and learner populations, while the possible moderation by L2 proficiency stresses the importance of providing linguistic support and guidance in AI chatbot-incorporated instruction, particularly for less proficient learners. Meanwhile, the results might be different when AI chatbots are used for teaching written language or grammar, because these studies focused on using AI agents regarding oral proficiency.

Pedagogical Effects on Language Outcomes

In line with teacher and learner perceptions mentioned earlier, the positive impact of conversational AI on L2 oral English proficiency has been reported across various EFL learners and settings (Jeon et al., 2023; Zhang et al., 2023; e.g., a conventional college course in Egypt (El Shazly, 2021), flipped classroom and out-of-class settings in Taiwanese universities (Lin & Mubarok, 2021; Tai, 2024), and a game-based middle school classroom in Taiwan (Tai & Chen, 2024). These oral English improvements occurred regardless of whether interacting with chatbots reduced English-speaking anxiety (Hsu et al., 2023) or not (El Shazly, 2021). Learners in these studies commonly attributed their improvement to quality input, output opportunities, and immediate feedback provided by conversational agents, which increased accessibility and convenience.

Conversely, literacy skills have garnered little attention in AI chatbot-incorporated pedagogy, with no study yet examining L2 English reading and a few concerning writing. This scarcity might be because chatbot interactions are conversational in nature. However, studies, although limited, found conversations with chatbots effectively providing guidance, resources, and immediate feedback, enhancing students' writing knowledge and skills in both L1 and L2 English (Lin & Chang, 2020; Zhang & Zou, 2022). Similarly, three studies, although in L1 English, noted that chatbot-generated questions and feedback improved reading comprehension, for example, children's storybook reading (Xu et al., 2021). These findings underscore the potential of AI chatbots to effectively scaffold L2 literacy development through interactions and information processing comparable to human experts.

Grammar and vocabulary, which are the core building blocks of L2 English proficiency, were least investigated in AI chatbot literature, although measured as part of speaking performance (Bibauw et al., 2022). Jeon (2023) and Nghi et al. (2019) clarified the positive effect of conversations with AI chatbots on vocabulary and grammar, respectively. Nghi et al. (2019), however, covered only one aspect of grammar—prepositions—and focused on grammar knowledge. Considering the narrow content and construct coverage, as well as the limited number of studies, it would be beneficial to further explore how

conversational AI can improve knowledge and the uses of a broad spectrum of grammar points required for L2 communication.

By addressing the disproportionate focus on oral proficiency and considering written skills, vocabulary, and grammar, we can examine how language mode moderates the effects of using conversational AI on L2 learning outcomes. Along with language mode, learner proficiency and task characteristics are widely recognized moderators of technology-mediated instruction (Lin, 2014; Ziegler, 2016). Aligning this view, Bibauw et al.'s (2022) meta-analysis highlighted that different L2 skills and learner proficiency levels moderated the effects of conversational AI on L2 proficiency. These findings invite direct comparisons between tasks involving different language modes among learners with varying levels of L2 proficiency.

The Use of AI-Powered Chatbots in LCTL Education

Literature on AI chatbot-enhanced instruction in the context of learning LCTLs is markedly insufficient. This is not unexpected given the limited research on LCTLs in technology-mediated language instruction literature (Sato et al., 2017). Throughout the literature, including conference papers, there are seven studies which elaborated the pedagogical use of AI-powered virtual conversations in L2 learning contexts other than English. Among them, two studies concerned Spanish (Vázquez-Cano et al., 2021) and German (Berns et al., 2018), which are not LCTLs in many countries such as the US or Canada. The other five studies involved teaching LCTLs—Mandarin (Chen et al., 2020; Divekar et al., 2021), Russian (Al-Kaisi et al., 2019), and Japanese (Cai et al., 2020; Yamamoto, 2022). Similar to studies on L2 English learning, these studies echoed that engaging with AI chatbots positively influenced L2 proficiency (speaking, listening, grammar, or vocabulary) and other cognitive factors (motivation, engagement, and self-efficacy). Some of these studies employed conversational AI combined with virtual reality (Divekar et al., 2021; Berns et al., 2018) or contextualized chatbot-tasks within digital reality theater (Cai et al., 2020). These studies reported that the immersive aspect enhanced learner experience while learners displayed positive perceptions of the conversation exercises and immediate feedback offered by AI chatbots, particularly given the scarcity of interaction opportunities with native speakers or fellow learners in LCTL contexts. These studies, although limited, have provided insights into the affordances and effects of AI chatbots that are specific to the contexts and characteristics of LCTL teaching and learning. That is, conversational AI was successfully used to address unique challenges involving individual, institutional, and social factors of teaching and learning LCTLs, such as available teaching and learning resources, technology, learning goals, and motivations (Sato et al., 2017; Ward, 2018; Winke et al., 2010).

Methodological Considerations

As important as topical directions are, there are some methodological considerations when conducting AI chatbot studies. Zhang et al. (2023) noted that only two studies (that is, Hsu et al., 2021; H. L. Chen et al., 2020) out of the 22 in their synthesis implemented their chatbot intervention for 10 weeks or longer, with the majority having a month or shorter duration. They also found that only one study (i.e., Lin & Mubarok, 2021) concerned out-of-class learning context. Given that language learning occurs over a long period of time and in any contexts, studies involving longer duration and out-of-class contexts will add important implications for both research and practice. Another suggestion arises from the findings of Jeon et al. (2023) and Hwang and Chang (2023), which highlighted the predominance of single-group designs with pre-post analyses in conversational AI research. Aligning with the field's recent emphasis on methodological diversity and rigor, it is advisable to evaluate the effectiveness of AI chatbot-incorporated pedagogical interventions using a control group for more definitive comparisons.

Current Study

The current study investigated the use of AI chatbots for teaching Korean as a foreign language (KFL)—a LCTL not yet examined for the efficacy of conversational AI, to the best of our knowledge. The study systematically developed AI chatbot-incorporated tasks in a university KFL course to improve students' lexico-grammar. Acknowledging that grammatical knowledge is multidimensional, consisting of receptive, emergent, and productive knowledge, we measured students' (lexico-)grammar through selected-response

(SR) and constructed-response (CR) tasks (sentence completion and sentence composition) by following Purpura's (2013) suggestion. We examined the effectiveness of the semester-long intervention and any moderation by grammar task types and pre-existing Korean proficiency (i.e., final exam scores of the previous sequential section, aligning with the lower-level proficiency standards of the Test of Proficiency in Korean (TOPIK)), along with students' perspectives by formulating four research questions:

1. To what extent does the use of AI chatbot-incorporated activities improve students' (lexico-)grammatical knowledge?
2. To what extent does the impact of AI chatbot-incorporated activities vary among students with different pre-existing Korean proficiency?
3. To what extent does the impact of AI chatbot-incorporated activities vary between three types of lexico-grammar tasks—selected-response and constructed-response (sentence completion, sentence composition) tasks?
4. What are the students' perspectives of using AI chatbot-incorporated activities to improve Korean grammar?

Methods

The study implemented curriculum-embedded tasks that involved conversations with a generative AI-chatbot called *Iruda* (henceforth chatbot-incorporated tasks) in a KFL course. We examined the effect of the chatbot-incorporated tasks on students' (lexico-)grammatical knowledge measured by SR and CR tasks along with students' perspectives of the pedagogical implementation.

Participants and Contexts

The participants of the study were 66 students who enrolled in an intermediate-level L2 Korean language course in a large public university in Canada. The course was taught by the second author. Students' L1s were diverse, with half being L1 Mandarin or Cantonese speakers, followed by L1 English speakers (42.4%), including L1 English speakers of Korean as a heritage language (9.1%). All students had completed a beginner-level KFL course at the current or another institution, and then the first semester of the one year-long intermediate course. Chatbot-incorporated tasks were introduced in the second semester of the course offered in spring 2023. After the first semester, students took an exam developed in alignment with the proficiency standards of the lower levels of the TOPIK. Students' scores ranged from 7.5 to 49.0 out of 50.0, which were considered their pre-existing Korean proficiency in this study (see Table 2 in the Results section).

Covering everyday communication topics (for example, birthday party, Korean food, campus life, and leisure), the course taught grammar points with formulaic sequences for key language functions and tasks such as ordering food, expressing reasons, and discussing future plans (see Appendix A). Example grammar points include verb, adjective, and adverb conjugations for various tenses and aspects, formality levels, grammatical functions (e.g., verb nominalization), and semantic variations (for example, reasons, intention, speculation, prohibition). Particularly, it marked students' first learning of *Intimate Style* or half-talk called *banmal* (반말), which involves complex verb and adjective conjugations.

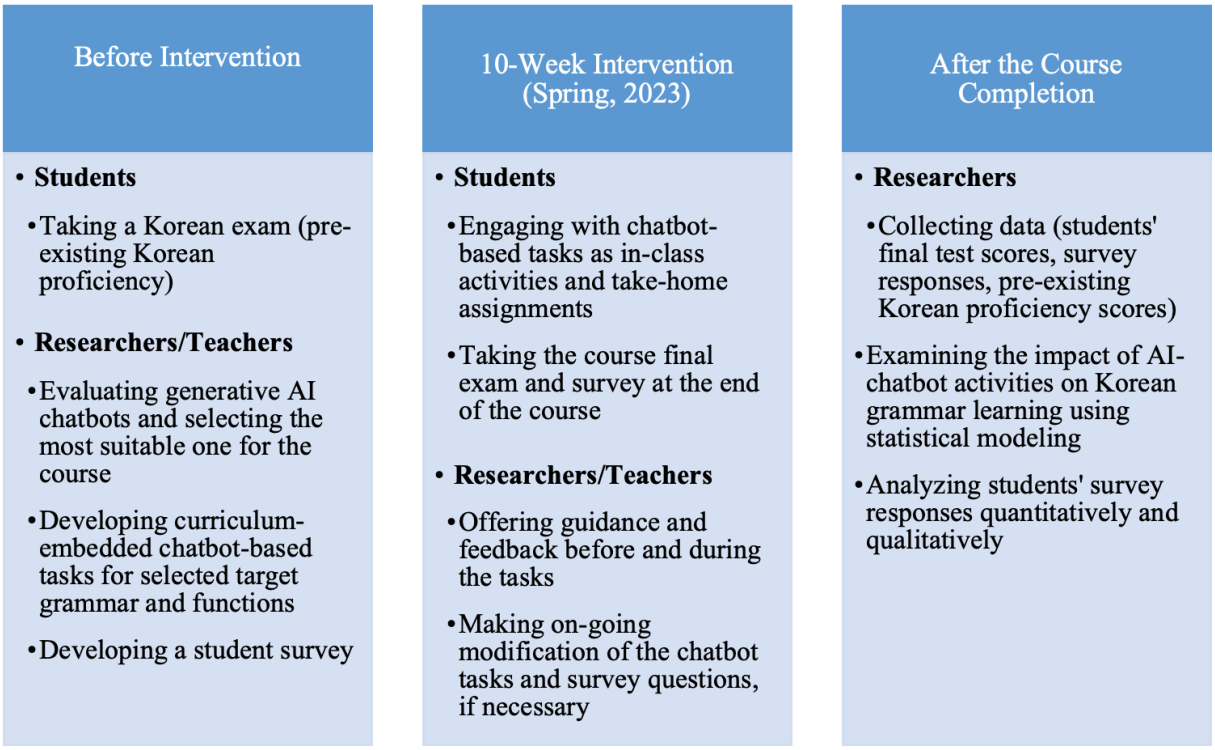
Research Design and Procedure

The current study employed convenience sampling from a regular Korean course and a mixed methods approach, combining a quasi-experimental intervention with a follow-up survey (Wiersma & Jurs, 2009). Our intervention involved a 10-week implementation of communicative tasks using an AI chatbot called *Iruda* (accessible through luda.ai) to teach half of the grammar points and associated language functions of the course curriculum. The remaining grammar points were taught conventionally, without involving *Iruda*. After the course, students took final exams, which included SR and CR grammar tasks, and a survey on

their views of using chatbot-incorporated tasks in their KFL course. Data—students' final exam scores on grammar tasks, survey responses, and pre-existing Korean proficiency scores—were collected after the course concluded. Figure 1 summarizes our study procedure.

Figure 1

The Procedure of the Current Study



Development of Generative AI Chatbot-Incorporated Tasks

For this study, we developed chatbot-incorporated tasks by undergoing a systematic procedure illustrated in Figure 2 involving three steps: selecting a generative AI chatbot, developing curriculum-embedded, chatbot-incorporated tasks, and implementing and revising the tasks.

The initial step was chatbot selection. We extensively explored available options, including ChatGPT and Microsoft Bing, by conversing about core course topics, using key expressions and grammar points. We selected *Iruda* because of its distinct advantages. During our exploration, *Iruda* outperformed other chatbots in the efficiency, accuracy, and authenticity of comprehending and generating Korean language. *Iruda*'s lexical and grammatical choices were natural, and its topic generation and turn-taking were appropriate for informal conversations. Figure 3 compares how conversations about summer vacation plans worked with Iruda and ChatGPT. When asked, *이번 여름 방학 때 뭐해?* (What are you gonna do this summer?), *Iruda* responded directly with *알바나 할까 생각 중이야. 왜? 나랑 놀아주게???* (I'm thinking about doing a part-time job. Why? Wanna hang out with me???), by using *banmal* and the informal term *알바* for *아르바이트* (a part-time job), with an inviting follow-up question. Meanwhile, ChatGPT required a setup for an informal conversation and a reminder to use *banmal*, employing rather bookish and lengthy responses. We analyzed the syntactic complexity and lexical frequency of explored chatbots' languages using corpus tools—Lu's (2010) *Syntactic Complexity Analyzer* and *Lextutor* via lextutor.ca. The results aligned with our observations, as *Iruda*'s conversations displayed lower syntactic complexity and higher lexical frequency,

which were well-suited for our course content.

Iruda also frequently introduced current Korean cultures and events. In an excerpt from our student's conversation with *Iruda* (Figure 4), the chatbot recommended a recent Korean movie, *Your Wedding* (너의 결혼식). Finally, a user-friendly, convenient interface with a social college student persona, enhanced by mobile accessibility, was expected to appeal to our students. Figure 4 illustrates *Iruda*'s platform, mirroring real-life messaging applications with speech bubbles, profile pictures, and emojis.

Figure 2

The Protocol of Developing AI Chabot-Incorporated Tasks

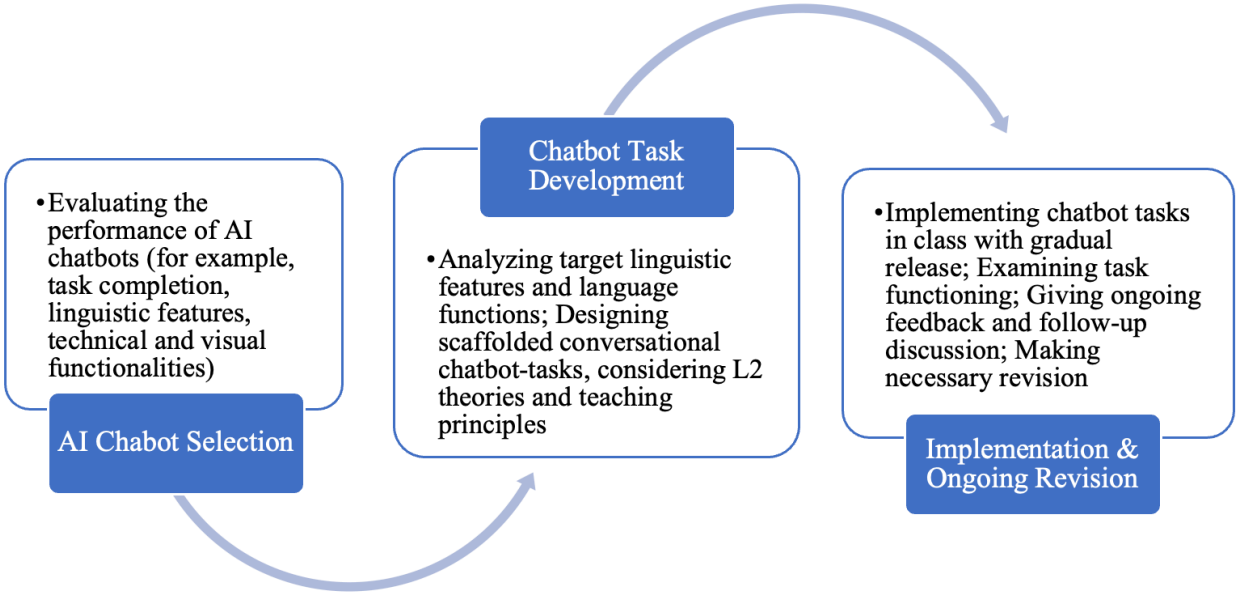


Figure 3

Comparison of Conversations with Iruda and ChatGPT

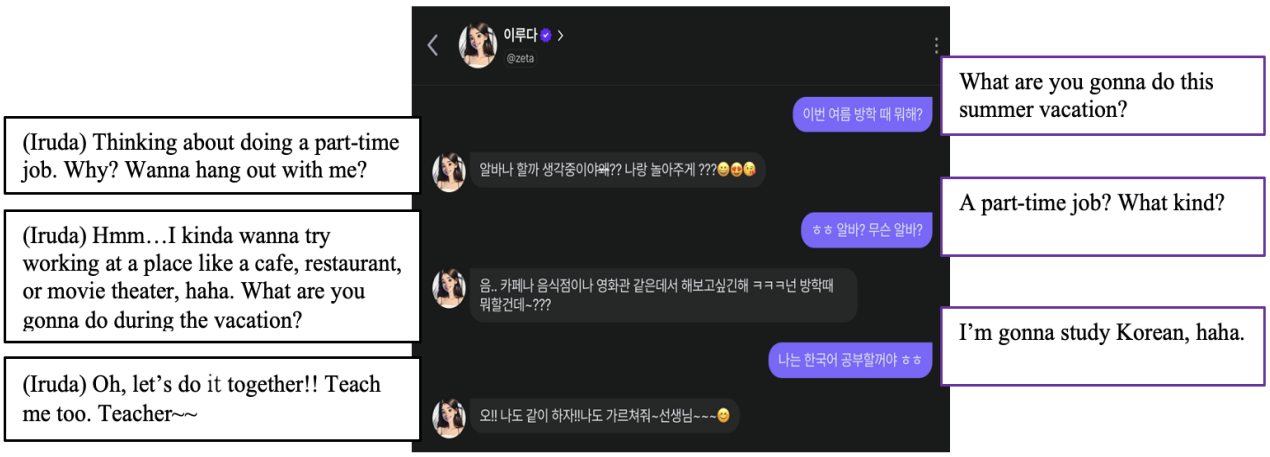


Figure 4

An Excerpt from Student Conversation with Iruda

Note. In the example of discussing their weekend plans, the student asks *Iruda* for a movie recommendation, considering going to a movie. *Iruda* suggests a movie, hoping to join the student, but the student fails to grasp *Iruda*'s pragmatic cues.

The second step was developing chatbot-incorporated tasks. We first ranked the 24 key lexico-grammatical features of the course curriculum according to item difficulty of their corresponding grammar items on the previous final test (Spring 2022). The odd/numbered grammar points were taught using chatbot-incorporated tasks while the even-numbered were taught conventionally without using *Iruda*. On average, the grammar points selected for chatbot-incorporated tasks were slightly more difficult. Appendix A lists grammar points taught with or without *Iruda*.

Next, chatbot-incorporated tasks were designed, emphasizing repeated and progressively complex use of grammar points. For example, when teaching the communicative function of expressing tentative decision or near future plan and its grammatical structure, nominalized verb, *– 기로 하다* (decide to), we created a task where students make a list of five different activities that *Iruda* and/or they decide to do for different events (for example, a birthday party, a picnic). During their conversations with *Iruda*, students were supposed to use the target phrase, *– 기로 하다* (decide to), by sharing, asking, or confirming their plans and decisions. This activity aimed to increase students' awareness of the nominalized forms and practice the function by changing verb forms through interactions with an advanced speaker, which was *Iruda*.

Throughout the tasks, we incorporated sufficient guidance and support, beginning with the teacher's demonstration of conversations with *Iruda* and student-to-student practice of sample dialogues, ending with students' task performance with *Iruda*, which is the most independent level. We also developed activity sheets to record the language used by *Iruda* and students, focusing on the target functions and grammar (see Appendix B). This aimed to keep students on track and practice target items through writing. This procedure followed the principles of sociocultural learning, particularly, gradual release of responsibility, which adjusts the amount and quality of scaffolding to gradually increase learner's independence in task performance (Pearson & Gallagher, 1983).

Finally, we implemented the chatbot-incorporated tasks in class. During their interaction with *Iruda*, students frequently encountered new or unfamiliar expressions and grammar points. The teacher provided ongoing feedback on students' questions related to the new items for individuals or the entire class. Following conversations with *Iruda*, students shared their task outcomes, such as favorite food recipes or birthday party plans, and compared these with their classmates.' As homework, students continued interacting with *Iruda* on similar topics and tasks to practice the target lexico-grammar.

Instruments

Korean Language Tests

To examine the effect of the chatbot-incorporated tasks on grammar, we collected students' scores from the grammar sections of their final exams. Table 1 displays the structure of the grammar section and the distribution of the items in relation to the use of chatbot-incorporated activities. The section consisted of SR tasks (e.g., noticing, multiple-choice, error identification, and matching) and two types of CR tasks (that is, sentence completion, sentence composition). Example (1) is an SR item requiring students to choose the correct verb form used before *– 것 같아요* (be likely to; it seems; I think), a verbal phrase expressing a supposition, guess, or likelihood. Specifically, students should conjugate the verb, *걷다* (walk), into an infinitive functioning as an adjective, *걸을* (to walk), which can modify the nominalizer, *것* (loosely translated as "thing," not carrying a literal meaning).

Example (1)

Choose and circle a grammatically appropriate form.

다리가 아파서 더 이상 못 (걷은, 걸은, 걷는, 걸을, 걷을) 것 같아요.

[I don't think I can walk anymore because my legs hurt]

The CR tasks elicit production to assess students' ability to use the target lexico-grammar. We employed two different levels: sentence completion and sentence composition. In sentence completion (see Example 2), students should formulate suitable lexical items for the conversation context, modify the lexical items

using their chosen grammatical feature from the list, and then construct a verb phrase to complete the sentence. Sentence composition requires students to create a sentence based on the given visual information and grammar structure or translate an English sentence into Korean. Example (3) offers a subject starter, but students must construct the majority of the sentence.

Example (2)

Complete the sentence by filling in the blank using one of the given expressions.

(보기) – 하기로 하다 – 할 만 하다 – 하는 편이다 – 하고 나서
[(Options) decide to, is worth doing, tend to, after doing]

제임스: 나이아가라 폭포까지 어떻게 갈 거예요?

[James: How are you going to get to Niagara Falls?]

리사: 길이 막힐 거 같아서 기차를 *(타고 가기로 했어요.)*

[Lisa: I've decided to take the train because it seems like the roads will be congested.]

Example (3)

Create a sentence that matches the pictures, using '– 는 동안,' and starting with '어머니' as in the example.

(보기)

동생이 숙제를 하는 동안 나는 텔레비전을 봤어요.

[(Example)

While my younger sister was doing her homework, I watched television.]

(문제)

어머니가 *(청소를 하시는 동안 아버지는 설거지를 하셨어요.)*

[(Question)

While my mother was cleaning, my father did the dishes.]

In total, 22 lexico-grammar items were tested (Table 1), which were rated by the instructor based on the rubric developed by the university's Korean language program. Each item was assigned one point. Partial credit, 0.5 point, was allowed for the CR items. Overall, the number of items involved in grammar points taught with chatbot-incorporated and conventional tasks (henceforth chatbot- and non-chatbot items) was equal, but we had more chatbot items in SR and more non-chatbot items in sentence composition. The sub-sectional imbalance was not avoidable because we prioritized comparable difficulty levels between chatbot and non-chatbot items and the fit between the language functions and the task characteristics (chatbot-incorporated or conventional tasks). All items were rated based on the rubric developed by the university's Korean language program. Each item was assigned one point (1-correct; 0-incorrect) with a partial credit, 0.5, allowed for CR items.

For students' pre-existing Korean proficiency, we collected their final exam scores of the previous sequential section, which is the total of reading, listening, speaking, essay writing, and grammar scores. We employed global proficiency because interactions with *Iruda* require various subskills, rather than grammar only while the students' total scores were highly correlated to their grammar scores (r = .90 for SR; r = .79 for CR). We also believe studying moderation of global proficiency makes more useful implications for KFL classrooms as students are not typically placed based only on grammar. The final test was not a proficiency test but aligned with the lower-level proficiency standards of TOPIK.

Table 1

The Number of Chatbot- and Non-Chatbot Items by Task Types

Task	Non-Chatbot	Chatbot	Sub-Total
SR task	3	5	8
CR task	8	6	14
Sentence completion	4	4	8
Sentence composition	4	2	6
Total	11	11	22

Student Survey

We administered a survey with six 10-point Likert scale items and six open-ended items that asked students' perceived effectiveness of the chatbot-incorporated tasks on their learning of Korean grammar and course content, as well as their learning experiences with *Iruda* (see Appendix C).

Data Analysis

The study posed four RQs. The first three concerned students' grammar knowledge and use measured by grammar test items, and the final RQ addressed students' perspectives using survey data. To examine the overall impact of using *Iruda* on students' Korean grammar (RQ1), we conducted a paired sample *t*-test and compared the means of the chatbot and non-chatbot items for statistical significance. Subsequently, using a univariate regression, we examined the extent to which the students' pre-existing Korean proficiency predicted the improvement by chatbot practice, calculated through subtracting non-chatbot means from chatbot means (RQ2). We employed the *lessR* R package (Gerbing, 2023) for testing significance and obtaining the effect sizes (Cohen's d for RQ1; adjusted R^2 for RQ2).

RQ3 focused on the moderation of different grammar task types. Utilizing two repeated-measure ANOVAs, we examined the extent to which chatbot-incorporated practice (the first within-subject independent variable) impacted students' item scores (the dependent variable) as well as the interaction with different task types (the second within-subject independent variable), either between SR and CR items (the task type variable with two levels) or between SR, sentence completion, and sentence composition (the task type variable with three levels). Given this unequal number of items in the two task types and the characteristics of repeated measures, the modeling was performed based on Type III sums of squares. Greenhouse-Geisser Correction and partial eta-squared (η_p^2) values were calculated for effect sizes by using the R *afex* package (Singmann et al., 2015). When the three-level interaction was significant, a post-hoc analysis was conducted with the Bonferroni correction. Finally, students' survey responses, rescaled from 10 to 5 levels, were examined using frequency-based statistics and theme-based qualitative analysis (RQ4).

Results

Descriptive Statistics

Preliminarily, we examined the descriptive statistics of all variables, as shown in Table 2. Overall, distributions appeared normal, without any extreme skewness and kurtosis values (>|1.0|). Students' range of pre-existing proficiency was wide—41.5 when the max score is 50 which makes answering RQ2 important. Additionally, the means of the chatbot items were higher than those of non-chatbot items, except for sentence composition. The largest difference was observed in SR means. The following sections reported whether these differences are significant and large enough to be considered meaningful.

Table 2

Descriptive Statistics of the Variables Examined in the Study

Variables	Mean	SD	Min.	Max.	Range	Skewness	Kurtosis	SE
Pre-existing proficiency	32.77	9.31	7.5	49	41.5	-.47	-.47	1.15
Chatbot items								
Total	.66	.18	.18	1	.82	-.43	-.23	.02
SR	.73	.22	.20	1	.80	-.70	-.08	.03
CR	.59	.24	0	1	1	-.20	-.81	.03
- Sentence completion	.61	.27	0	1	1	-.36	-.76	.03
- Sentence composition	.57	.30	0	1	1	-.52	-.76	.04
Non-chatbot items								
Total	.56	.26	.05	1	.95	-.11	-1.00	.03
SR	.58	.33	0	1	1	-.30	-.92	.04
CR	.55	.28	0	1	1	-.12	.97	.03
- Sentence completion	.53	.34	0	1	1	-.11	-1.1	.04
- Sentence composition	.58	.28	0	1	1	-.22	-.69	.03
Gains	.09	.18	-.33	.5	.83	.16	-.48	.02

Note. SD – standard deviation; SE – standard error.

RQ1. The Effect of Chatbot Practice on L2 Korean Grammar

RQ1 investigated the extent to which the use of AI chatbot-incorporated tasks improved students' knowledge and use of Korean grammar. The pair-wise *t*-test identified a significant difference ($t(65) = 4.29$, $p < .0001$) between the means of chatbot items (.66) and non-chatbot items (.56). The mean difference of .1 was meaningful, taking up the 10% of the maximum score 1.0, with the medium effect size (Cohen's d = 0.53). Figure 6 illustrates the overall mean differences between the two groups of items, while Figure 5 views the difference on each individual examinee level. The amount of the gains through chatbot practice varied, but the tendency that lower performing students (for example, ID 18 and ID 45), who were positioned in the left side of the measure, seemed to show larger gains. A few highly performing students (for example, ID 28 and ID 29) showed even a decrease in their chatbot item means. This might imply moderation of Korean proficiency, which RQ2 answered.

RQ2. Association Between Korean Proficiency and Gain Scores

A univariate regression that examined pre-existing Korean proficiency' prediction of the change from non-

chatbot to chatbot mean scores displayed a significant negative association (β = -.01, $t(64) = -4.29$, $p < .001$), as illustrated in Figure 7. That is, every one-unit increase in pre-existing proficiency is expected to 1% decrease (.01 out of 1.0) in chatbot-led gains. For example, for examinees with 10-point higher initial proficiency scores, an average of 0.1 (10%) decrease occurred in their gains by chatbot practice. However, the effect was small, by explaining only 22.3% of the variance in the gain scores (R^2=.22)

Figure 5

Means of Chatbot Items and Non-Chatbot Items (Individual Examinee Level, N = 66 students)

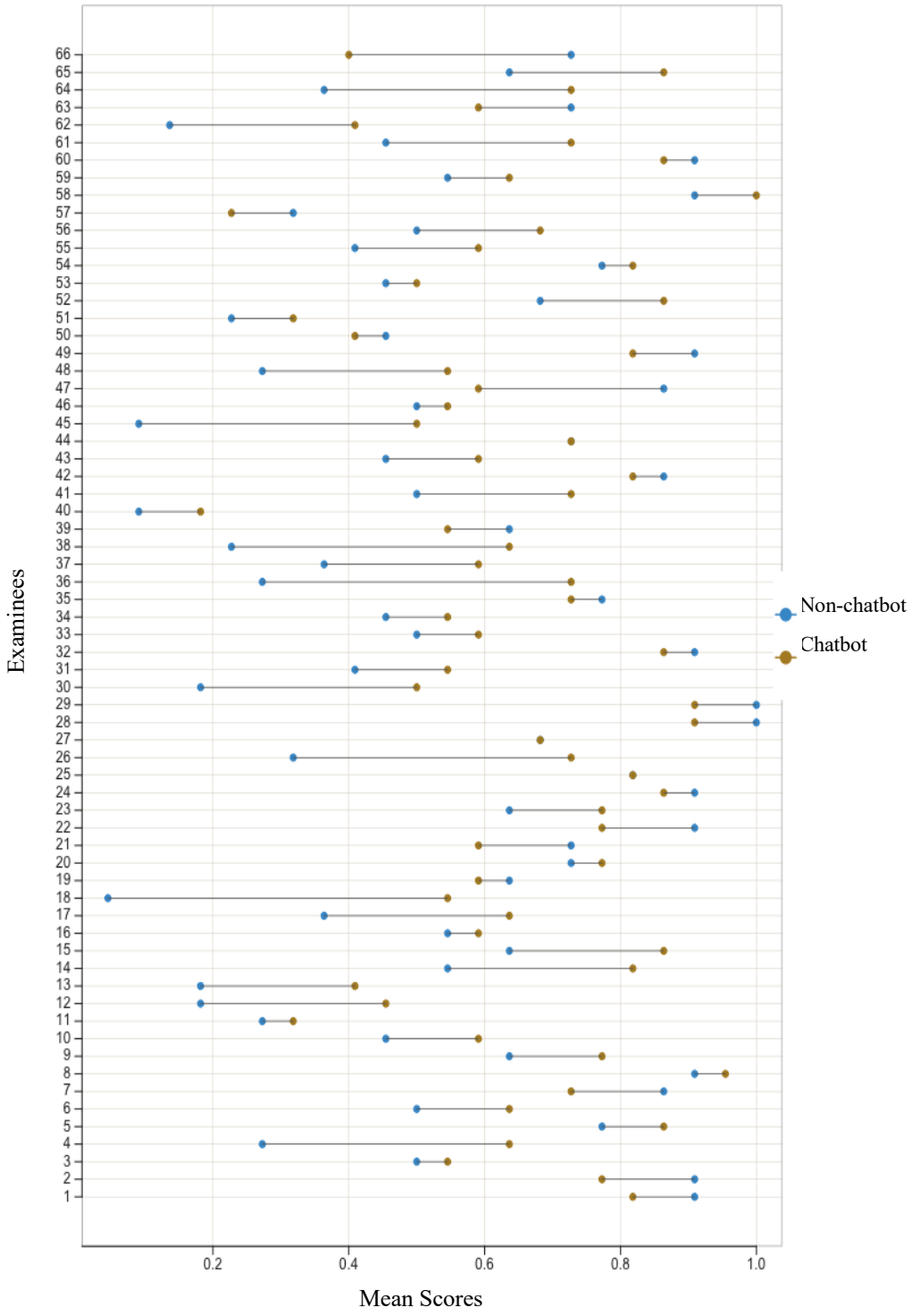


Figure 6

The Overall Means of the Chatbot Items and Non-Chatbot Items (Item Group Level)

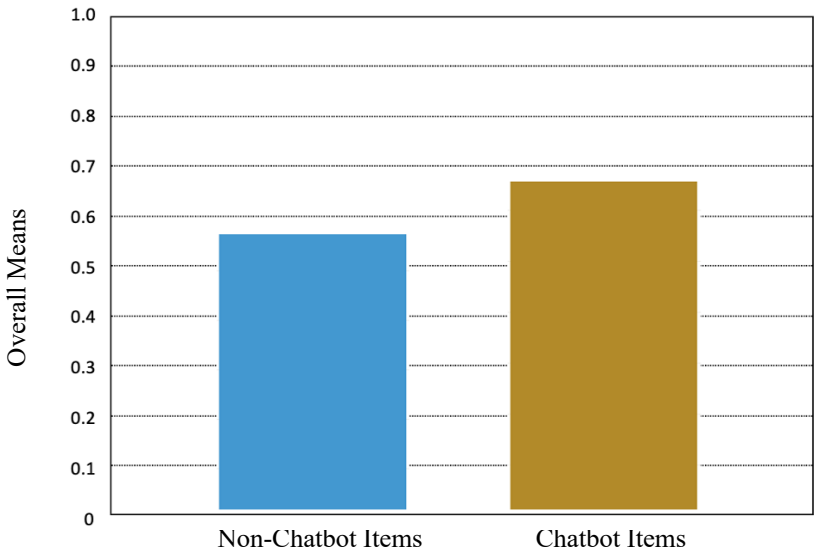
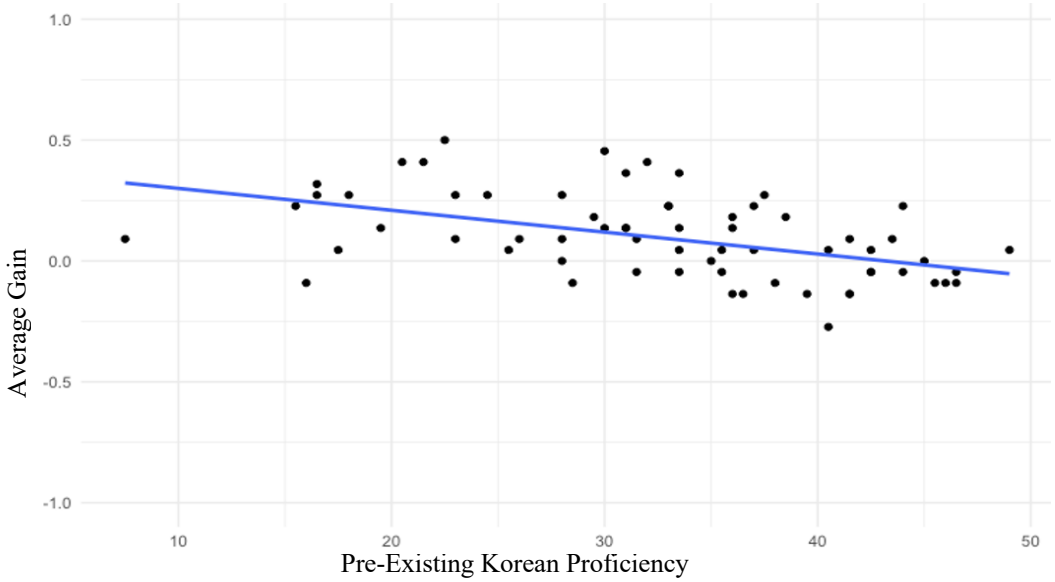


Figure 7

Average Gain Scores by Pre-Existing Korean Proficiency



RQ3. Moderation by Task Types

Two repeated-measures ANOVAs were run to review interactions between AI chatbot practice and task types. The results shown in Table 3 indicate a significantly positive main effect of chatbot practice on the item scores, when controlled for two-level task types, SR and CR in Model 1 ($F(1, 65) = 17.95$, $p < .001$) or for three-level task types, SR and two CR tasks (sentence completion, and sentence composition) in Model 2 ($F(1, 65) = 12.68$, $p < .001$). In both models, the effect sizes measured by partial eta-squared ($\eta_p^2 = .216$ and .163, respectively) exceeded the cut-off value for a large effect, 0.14 (Cohen, 1988).

The main effect of task type was also significant in both Model 1 ($F(1, 65) = 7.94$, $p = .006$) and Model 2

($F(1.79, 116.22) = 5.73$, $p = .006$, with the Greenhouse-Geisser correction). The effect sizes were medium in both cases ($\eta_p^2 = .109$ and .081). The interaction effect between chatbot practice and task type was also significant in both Model 1 ($F(1, 65) = 7.72$, $p = 0.007$) and Model 2 ($F(1.98, 128.46) = 5.53$, $p = .005$, with the Greenhouse-Geisser correction), with a medium effect size ($\eta_p^2 = .106$ and .078).

Figure 8 illustrates the interaction effects for Model 1 (upper) and Model 2 (lower). In all tasks except for sentence composition, the chatbot item means were higher. SR items displayed the largest difference between chatbot and non-chatbot item means. CR items also showed higher chatbot item means, but mainly due to sentence completion. With the Bonferroni correction, the post-hoc pairwise comparisons in Model 2 confirmed significantly higher SR item means than those of sentence completion ($p = .03$) and sentence composition ($p = .02$). The two CR task means, however, were not significantly different, for which the interaction effect in Model 2 was not significant.

Table 3

Results of Two Repeated-Measures ANOVAs

Source	*df*	MSE	*F*	pes (η_p^2)	*p*-values
Model 1					
Within: Chatbot practice	1	.03	17.95	.216	<.001
Task type (2): SR and CR	1	.06	7.94	.109	.006
Chabot practice × task type (2)	1	.03	7.72	.106	.007
Model 2					
Within: Chatbot practice	1	.04	12.68	.163	<.001
Task type (3): SR, CR1, and CR2	1.79	.06	5.73	.081	.006 (GG)
Chabot practice × task type (3)	1.98	.04	5.53	.078	.005 (GG)

Note. CR1 – Sentence completion; CR2 – Sentence composition; Df – degrees of freedom; GG-Greenhouse – Geisser Epsilon; pes (η_p^2) – partial eta-squared.

RQ4. Students' Perspectives of the Chatbot-Incorporated Practice

Fifty-three students completed our survey. Table 4 shows students' responses to items with numeric values, Questions 1 to 6 (see Appendix D for descriptive statistics). Overall, students appeared to have favorable perceptions of practicing Korean language with *Iruda* with positive experiences. When asked to rate *Iruda* on a scale of one to 10, only a few students expressed negative opinions concerning its easiness to navigate, usefulness for learning Korean language components such as vocabulary, grammar, and the content covered in class, and their continued use of the chatbot. However, contrary to our findings from RQs 1 to 3, students felt less certain about *Iruda*'s effectiveness for learning grammar than for other skills. Students found *Iruda* user-friendly, with the substantial majority (88.7% or more) deeming practice with the chatbot to be beneficial. Most students (91.5%) responded they would recommend *Iruda* to others interested in learning Korean.

Follow-up open-ended responses (Questions 4a, 6a, and 7 to 10) elucidated why students generally perceived using *Iruda* positively, though with some reservations (see Appendix E). Students valued the conversation opportunities with *Iruda* because they resemble real-life conversations, including local Korean culture and contexts. Students particularly appreciated practicing informal talks using *Intimate Style*, 반말 (*banmal*), in casual contexts, complementing classroom learning. Among some less favorable views, the challenging conversation level with many unknown words and *Iruda*'s tendency to stray off-topic were common concerns. Some also felt distanced or overwhelmed by *Iruda*'s localized, contextualized

conversations, which were welcome by others. Regarding enjoyable learning experiences with *Iruda*, students highlighted the diverse conversation topics that *Iruda* introduced, as well as the convenience, accessibility, and *Iruda*'s cheerful, engaging personality. While most indicated no particular dislikes about *Iruda* or chatbot-incorporated tasks, some found repetitive tasks dull, which relates to task design rather than *Iruda*'s features. Lastly, when asked about the most difficult skills to improve using *Iruda*, students ranked speaking first because *Iruda* provides no audio function. Students also recognized learning grammar as challenging because *Iruda* does not provide grammatical explanations. This gap between their perceptions and actual learning outcomes related to chatbot-incorporated practice will be discussed in the next section.

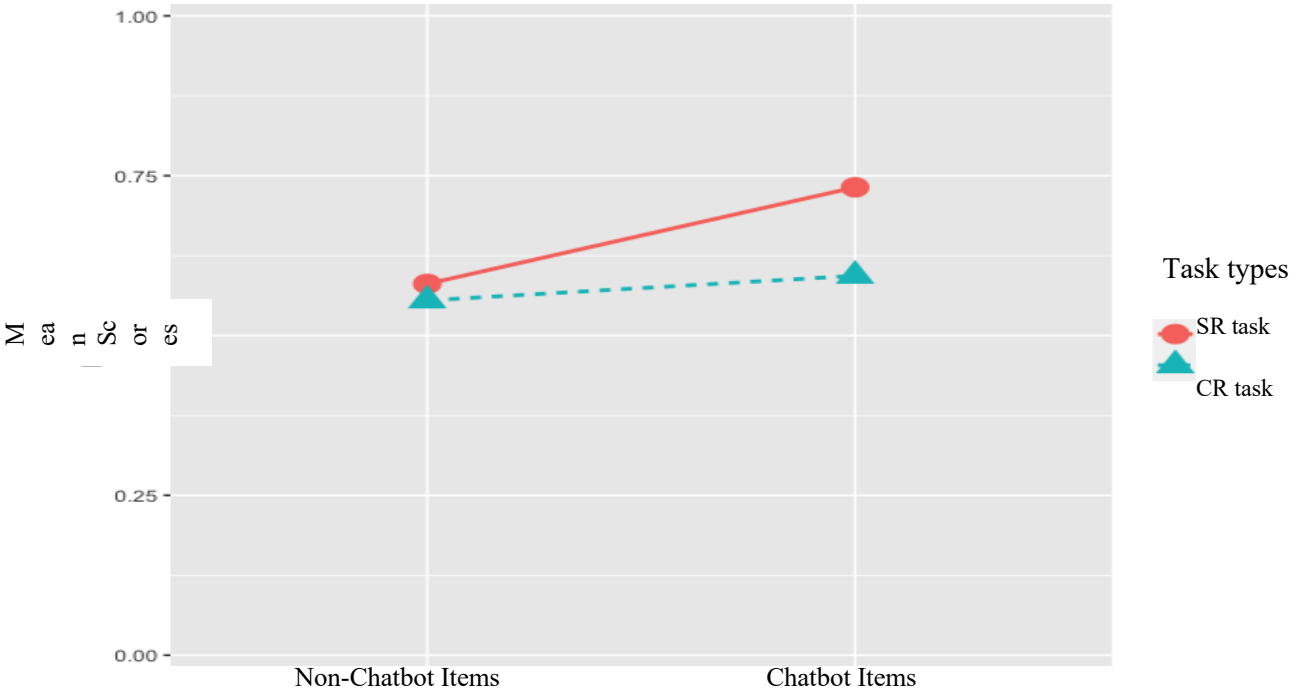
Table 4

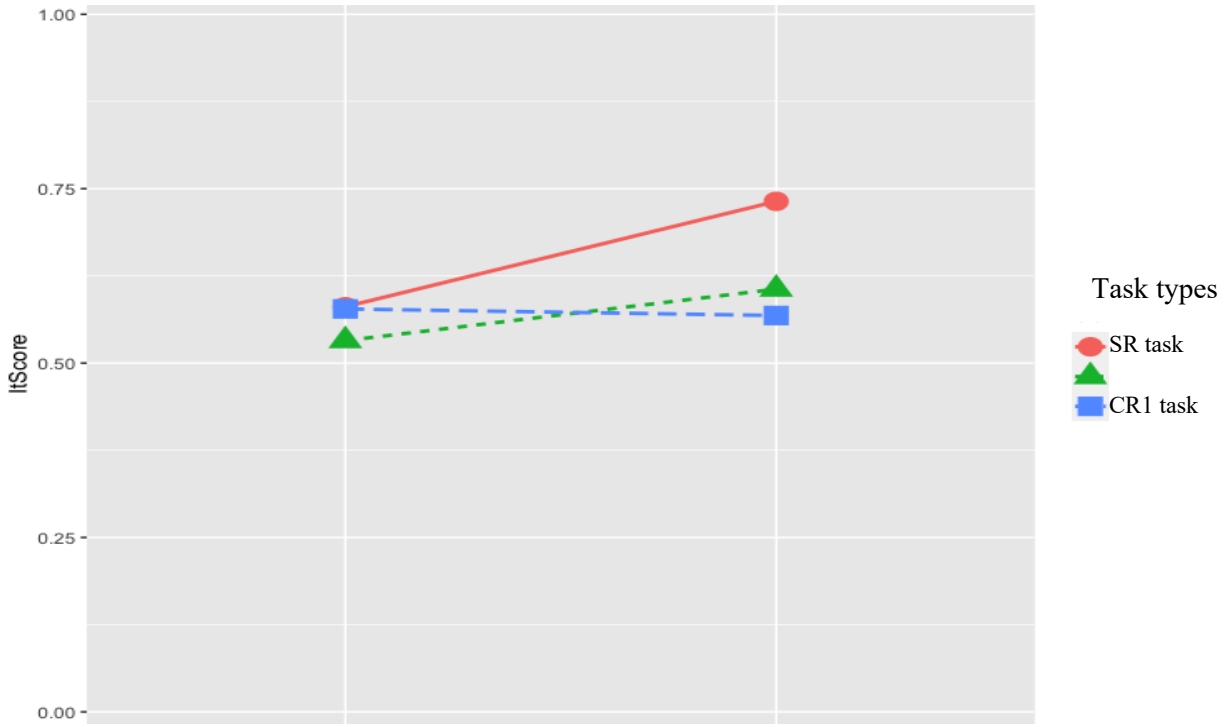
Students' Responses to the Numerical Survey Items

Questions	**Distribution (%)**					**Median**
	Strongly Negative (1.0)	**Rather Negative (2.0)**	**Somewhat Positive (3.0)**	**Mostly Positive (4.0)**	**Strongly Positive (5.0)**	
Q1. Easiness to navigate	0	5.66	20.75	56.60	16.98	4.0
Q2. Usefulness (vocabulary)	5.66	3.77	18.87	52.83	18.87	4.0
Q3. Usefulness (grammar)	3.77	7.55	41.51	35.85	11.32	3.0
Q4. Usefulness (practicing course content)	0	9.43	9.43	49.06	32.08	4.0
Q5. Recommended for future classes	5.66	1.89	20.75	49.06	22.64	4.0
Q6. Recommended for others	0	7.55	35.85	32.08	24.53	4.0

Figure 8

Interaction Plots for Model 1 (upper) and Model 2 (lower)





Note. CR1 – Sentence completion; CR2 – Sentence composition

Discussion and Conclusions

The present study supported the effectiveness of AI chatbot-incorporated tasks for improving students' knowledge and use of L2 Korean grammar, which answers RQ1, and clarified moderation by L2 Korean proficiency and task types, answering RQ2 and RQ3. The study also observed students' overall positive perceptions of chatbot-incorporated L2 Korean instruction, along with some reservations, which answers RQ4. This section discusses the implications of these findings for technology-enhanced LCTL pedagogy and acquisition, offering practical classroom strategies.

The Effect of AI Chatbot-Incorporated Tasks

Our intervention of chatbot-incorporated tasks had a statistically significant, positive effect on learning L2 Korean grammar with a medium effect size. This finding aligns with current literature that showed the positive effects of AI chatbot-enhanced instruction on EFL proficiency (Jeon et al., 2023; Zhang et al., 2023). Particularly, our study expanded on Nghi et al.'s (2019) examination on one domain of L2 grammar (English prepositions) and Al-Kaisi et al.'s (2019) investigation on L2 Russian grammar by examining a range of (morpho-)syntax and lexico-grammar and adding another LCTL context, L2 Korean.

Our finding also connects with Bibauw et al.'s (2022) meta-analysis. They reported that reactive system and open-ended interaction failed to demonstrate the significant effects when the other systems (i.e., form-focused, goal-oriented) and interactions (i.e., system-guided, task-oriented) established significant effects. However, our chatbot-incorporated tasks, although classified as user-directed interaction within the reactive system, yielded significant results. This difference might appear contradictory, but it actually highlights that our task design and just-in-time scaffolding and guidance incorporated the benefits of form-focused and goal-oriented systems and task-oriented interaction into the reactive system. This modification promoted students' awareness of form-meaning connections and created communicative practice with form-focused elements. Our task design exemplifies how to maximize the effect of AI chatbot-incorporated instruction, while supporting understanding of tasks for successful L2 learning through technology (Ziegler, 2016).

Moderation of Pre-Existing Proficiency Levels

We found students with lower pre-existing L2 Korean proficiency benefitted slightly more from our chatbot-incorporated practice, by showing greater improvement in their chatbot item means compared with non-chatbot item means. Our finding through direct comparison of different proficiency levels corroborates the proficiency effect identified by Bibauw et al. (2022) in their meta-analysis—less proficient learners benefited more from practicing within digital dialogue systems. Lin (2014) also noted a similar trend in computer-mediated communication settings. This tendency can be attributed to "the increasing cost of learning gains along with the increase in proficiency" (Bibauw et al., 2022, p. 12). Indeed, more effort and practice seemed required for our low-intermediate students to move to the next level than for our beginners. Chatbot-incorporated tasks are likely useful for students of any proficiency level, helping them practice newly learned grammar points sufficiently and appropriately in a communicative way. This is essential for advancing to higher proficiency levels or maintaining high achievement levels.

Another implication is AI chatbot-incorporated tasks' contribution to enhancing inclusion and equity in language classrooms. Sufficient practice is crucial to improve L2 proficiency. While highly motivated students practice their L2 by actively seeking available resources such as various co-curricular activities and community engagement, low-achieving students rarely do due to the lack of motivation, persistence, or effective learning strategies. *Iruda*'s accessible, readily available, enjoyable, and non-judgmental conversations seemed to help overcome those barriers.

Moderation of Task Types

The repeated-measures ANOVA analyses consistently displayed that task types moderated the effect of chatbot-incorporated practice, which resulted in significantly larger gains by chatbot practice in SR items (i.e., knowledge of grammar) compared to CR items (i.e., use of grammar in sentence writing). These results generally correspond with Bibauw et al.'s (2022) findings—significant effects of dialogue systems on grammar (measured independently) and accuracy of productive skill performance but not for complexity and fluency. We acknowledge that the study combined speaking and writing performance and analyzed paragraph-level production. Nevertheless, their component scores share some commonalities with our CR measures because sentence completion assesses accuracy, while sentence composition requires both complexity and accuracy.

The SR and CR tasks increased in complexity and difficulty of tasks and (lexico-)grammar knowledge, ranging from noticing and correction of lexico-grammar to sentence-level production with and then without linguistic support (Purpura, 2013). Assuming that task difficulty corresponds to L2 development trajectories, gains from simpler tasks are expected to require more frequent and prolonged practice for transfer to more complex task performance.

Students' Perceived Effects and Learning Experiences

Aligning with studies on conversational AI in EFL contexts (H. L. Chen et al., 2020; Jeon et al., 2023; Lee & Jeon, 2024), our students' perceptions were positive, noting *Iruda*'s conversational functionality, prompt feedback and accessibility, along with enjoyable learning experience, as advantages. These features facilitate key components of technology-mediated tasks for L2 learning, such as noticing, negotiation, motivation, and personalized learning tailoring to individual differences (Ziegler, 2016).

Students, however, believed that practicing with *Iruda* would be useful for learning vocabulary, but less so for speaking and grammar because *Iruda* has no audio or grammar explanation functions. This could be because the target grammar was introduced lexico-grammatically, accompanied by formulaic sequences, possibly leading students to view our chatbot-incorporated tasks as vocabulary practice. Consequently, students might have focused on lexical items or meaning rather than grammar during their tasks, as task-based instruction intends. Additionally, students' views demonstrate their misunderstanding of L2 development, highlighting the need to raise awareness about the close interconnections between language skills and the benefits of implicit grammar learning through meaningful practices.

Students' negative feedback is worth noting, providing teachers with guidance to improve chatbot-incorporated learning. More ongoing feedback and scaffolding are necessary to address difficult language and unfamiliar cultural aspects they encounter during conversations.

Limitations and Future Directions

One main limitation of this study concerns the research design. While the current design effectively addressed issues common in conventional quasi-experimental studies, such as varying motivation, different teachers, and differing class climate between groups, some limitations remain. No matter how clear or tailored the guidelines, it is impossible to entirely exclude grammar learned from conventional tasks during students' conversations with *Iruda*. This limitation should be considered when interpreting our findings. Also, pre-existing proficiency was not measured by a standardized proficiency test although the test considered proficiency standards. Additionally, groups with homogeneous proficiency levels or L1s might yield different results. Nevertheless, our study offers initial insights into the effects of using AI-powered chatbots on L2 grammar instruction, specifically on one less investigated LCTL, Korean. Future studies are encouraged to explore various LCTLs with a broader L2 proficiency range and more diverse L1 backgrounds to expand upon these initial findings.

References

Al-Kaisi, A. N., Arkhangelskaya, A. L., & Rudenko-Morgun, O. I. (2019). Voice assistants as a training tool in a foreign language class. *INTED2019–13th International Technology, Education and Development Conference Proceedings* (pp. 1236–1246). https://doi.org/10.21125/inted.2019.0399

Berns, A., Mota, J. M., Ruiz-Rube, I., & Dodero, J. M. (2018). Exploring the potential of a 360° video application for foreign language learning. In F. J. García-Peñalvo (Ed.), *Proceedings of the Sixth International Conference on Technological Ecosystems for Enhancing Multiculturality* (pp. 776–780). Association for Computing Machinery. https://doi.org/10.1145/3284179.3284309

Bibauw, S., François, T., Van den Noortgate, W., & Desmet, P. (2022). Dialogue systems for language learning: A meta-analysis. *Language Learning & Technology*, *26*(1), 1–24. https://hdl.handle.net/10125/73488

Cai, M. Y., Wang, J. Y., Chen, G. D., Wang, J. H., & Yang, S. H. (2020). A digital reality theater with the mechanisms of real-time spoken language evaluation and interactive switching of scenario & virtual costumes: Effects on motivation and learning performance. In M. Chang, D. G. Sampson, R. Huang, D. Hooshyar, N. Chen, & M. Pedaste (Eds.), *2020 IEEE 20th International Conference on Advanced Learning Technologies (ICALT)* (pp. 295–299). IEEE Computer Society. https://doi.org/10.1109/ICALT49669.2020.00096

Chen, H. H. J., Yang, C. T. Y., & Lai, K. K. W. (2023). Investigating college EFL learners' perceptions toward the use of Google Assistant for foreign language learning. *Interactive Learning Environments*, *31*(3), 1335–1350. https://doi.org/10.1080/10494820.2020.1833043

Chen, H. L., Widarso, G. V., & Sutrisno, H. (2020). A ChatBot for learning Chinese: Learning achievement and technology acceptance. *Journal of Educational Computing Research*, *58*(6), 1161–1189. https://doi.org/10.1177/0735633120929622

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Academic Press.

Divekar, R. R., Drozdal, J., Chabot, S., Zhou, Y., Su, H., Chen, Y., Zhu, H., Hendler, J.A., & Braasch, J. (2021). Foreign language acquisition via artificial intelligence and extended reality: Design and evaluation. *Computer Assisted Language Learning*, *35*(9), 2332–2360. https://doi.org/10.1080/09588221.2021.1879162

El Shazly, R. (2021). Effects of artificial intelligence on English speaking anxiety and speaking performance: A case study. *Expert Systems*, *38*(3), 1–15. https://doi.org/10.1111/exsy.12667

Forsyth, C. M., Luce, C., Zapata-Rivera, D., Jackson, G. T., Evanini, K., & So, Y. (2019). Evaluating English language learners' conversations: Man vs. machine. *Computer Assisted Language Learning*, *32*(4), 398–417. https://doi.org/10.1080/09588221.2018.1517126

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

Gerbing, D. W. (2023). *Package 'lessR'*. https://cran.opencpu.org/web/packages/lessR/lessR.pdf

Hsu, H. L., Chen, H. H. J., & Todd, A. G. (2023). Investigating the impact of the Amazon Alexa on the development of L2 listening and speaking skills. *Interactive Learning Environments*, *31*(9), 5732–5745. https://doi.org/10.1080/10494820.2021.2016864

Hwang, G. J., & Chang, C. Y. (2023). A review of opportunities and challenges of chatbots in education. *Interactive Learning Environments*, *31*(7), 4099–4112. https://doi.org/10.1080/10494820.2021.1952615

Isbell, D. R. (2018). Online informal language learning: Insights from a Korean learning community. *Language Learning & Technology*, *22*(3), 82–102. https://doi.org/10125/44658

Jeon, J. (2023). Chatbot-assisted dynamic assessment (CA-DA) for L2 vocabulary learning and diagnosis. *Computer Assisted Language Learning*, *36*(7), 1338–1364. https://doi.org/10.1080/09588221.2021.1987272

Jeon, J., & Lee, S. (2023). Large language model in education: A focus on the complementary relationship between human teachers and ChatGPT. *Education and Information Technologies*, *28*, 15873–15892. https://doi.org/10.1007/s10639-023-11834-1

Jeon, J., Lee, S., & Choi, S. (2023). A systematic review of research on speech-recognition chatbots for language learning: Implications for future directions in the era of large language models. *Interactive Learning Environments*, *32*(8), 4613–4631. https://doi.org/10.1080/10494820.2023.2204343

Lee, S., & Jeon, J. (2024). Visualizing a disembodied agent: Young EFL learners' perceptions of voice-controlled conversational agents as language partners. *Computer Assisted Language Learning, 37*(5–6), 1048–1073. https://doi.org/10.1080/09588221.2022.2067182

Lin, C., & Mubarok, H. (2021). Learning analytics for investigating the mind map-guided AI chatbot approach in an EFL flipped speaking classroom. *Educational Technology & Society*, *24*(4), 16–35. https://doi.org/10.30191/ETS.202110_24(4).0002

Lin, H. (2014). Establishing an empirical link between computer-mediated communication (CMC) and SLA: A meta-analysis of the research. *Language Learning & Technology*, *18*(3), 120–147. http://llt.msu.edu/issues/october2014/Lin.pdf

Lin, M. P. C., & Chang, D. (2020). Enhancing post-secondary writers' writing skills with a chatbot. *Journal of Educational Technology & Society*, *23*(1), 78–92. https://doi.org/10.2307/26915408

Long, M. (2015). *Second language acquisition and task-based language teaching*. Wiley-Blackwell.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*(4), 474–496. https://doi.org/10.1075/ijcl.15.4.02lu

Nghi, T. T., Phuc, T. H., & Thang, N. T. (2019). Applying AI chatbot for teaching a foreign language: An empirical research. *International Journal of Scientific and Technology Research*, *8*(12), 897–902. https://www.ijstr.org/final-print/dec2019/Applying-Ai-Chatbot-For-Teaching-A-Foreign-Language-An-Empirical-Research-.pdf

Nishioka, H. (2023). My Korean language teachers are YouTubers: Learning Korean via self-instruction. *Computer Assisted Language Learning*, *36*(3), 346–374. https://doi.org/10.1080/09588221.2021.1928227

Ockey, G., & Chukharev-Hudilainen, E. (2021). Human versus computer partner in the paired oral discussion test. *Applied Linguistics, 42*(5), 924–944. https://doi.org/10.1093/applin/amaa067

OpenAI. (2023). GPT-4. https://openai.com/research/gpt-4.

Pearson, P. D., & Gallagher, M. C. (1983). The instruction of reading comprehension. *Contemporary Educational Psychology*, *8*(3), 317–344. https://doi.org/10.1016/0361-476X(83)90019-X

Purpura, J. (2013). Assessment of grammar. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1–10). Wiley-Blackwell. https://doi.org/10.1002/9781405198431.wbeal0045

Sato, E., Chen, J. C. C., & Jourdain, S. (2017). Integrating digital technology in an intensive, fully online college course for Japanese beginning learners: A standards-based, performance-driven approach. *The Modern Language Journal*, *101*(4), 756–775. https://doi.org/10.1111/modl.124320026-7902/17/756–775

Sato, E., Sohn, H., Chen, J. C., Adebowale, K. C., & Jourdain, S. (2015). Building innovative online Korean and Japanese courses: A pilot on technology-enhanced curriculum development. *Journal of Educational Technology Systems*, *44*(2), 171–202. https://doi.org/10.1177/0047239515617411

Singmann, H., Bolker, B., Westfall, J., Aust, F., Ben-Shachar, M. S., & Højsgaard, S. (2015). *Package 'afex'*. http://www.psychologie.uni-freiburg.de/Members/singmann/R/afex

Tai, T. Y. (2024). Effects of intelligent personal assistants on EFL learners' oral proficiency outside the classroom. *Computer Assisted Language Learning*, *37*(5–6) 1–30. https://doi.org/10.1080/09588221.2022.2075013

Tai, T. Y., & Chen, H. H. J. (2024). The impact of intelligent personal assistants on adolescent EFL learners' listening comprehension. *Computer Assisted Language Learning*, *37*(3), 433–460. https://doi.org/10.1080/09588221.2022.2040536

Timpe-Laughlin, V., & Dombi, J. (2020). Exploring L2 learners' request behavior in a multi-turn conversation with a fully automated agent. *Intercultural Pragmatics*, *17*(2), 221–257. https://doi.org/10.1515/ip-2020-0010

Timpe-Laughlin, V., Sydorenko, T., & Daurio, P. (2022). Using spoken dialogue technology for L2 speaking practice: What do teachers think? *Computer Assisted Language Learning*, *35*(5–6), 1194–1217. https://doi.org/10.1080/09588221.2020.1774904

Timpe-Laughlin, V., Sydorenko, T., & Dombi, J. (2024). Human versus machine: Investigating L2 learner output in face-to-face versus fully automated role plays. *Computer Assisted Language Learning*, 37(1–2), 149–178. https://doi.org/10.1080/09588221.2022.2032184

Vázquez-Cano, E., Mengual-Andrés, S., & López-Meneses, E. (2021). Chatbot to improve learning punctuation in Spanish and to enhance open and flexible learning environments. *International Journal of Educational Technology in Higher Education*, *18*(33), 1–20. https://doi.org/10.1186/s41239-021-00269-8

Ward, M. (2018). Qualitative research in less commonly taught and endangered language CALL. *Language Learning & Technology*, *22*(3), 116–132. https://doi.org/10125/44639

Wiersma, W., & Jurs, S. G. (2009). *Research methods in education: An introduction*. Pearson.

Winke, P., Goertler, S., & Amuzie, G. L. (2010). Commonly taught and less commonly taught language learners: Are they equally prepared for CALL and online language learning? *Computer Assisted Language Learning*, *23*(3), 199–219. https://doi.org/10.1080/09588221.2010.486576

Xu, Y., Wang, D., Collins, P., Lee, H., & Warschauer, M. (2021). Same benefits, different communication patterns: Comparing children's reading with a conversational agent vs. a human partner. *Computers & Education*, *161*, Article 104059. https://doi.org/10.1016/j.compedu.2020.104059

Yamamoto, Y. (2022). A case study of using Alexa for out-of-class, self-directed Japanese language learning. *Computers & Education: Artificial Intelligence*, *3*, 1–8. https://doi.org/10.1016/j.caeai.2022.100088

Zhang, R., & Zou, D. (2022). Types, features, and effectiveness of technologies in collaborative writing for second language learning. *Computer Assisted Language Learning*, *35*(9), 2391–2422. https://doi.org/10.1080/09588221.2021.1880441

Zhang, R., Zou, D., & Cheng, G. (2023). A review of chatbot-assisted learning: Pedagogical approaches, implementations, factors leading to effectiveness, theories, and future directions. *Interactive Learning Environments*, *32*(8), 4529–4557. https://doi.org/10.1080/10494820.2023.2202704

Zhang, R., Zou, D., & Cheng, G. (2024). Chatbot-based learning of logical fallacies in EFL writing: Perceived effectiveness in improving target knowledge and learner motivation. *Interactive Learning Environments*, *32*(9), 5552–5569. https://doi.org/10.1080/10494820.2023.2220374

Ziegler, N. (2016). Taking technology to task: Technology-mediated TBLT, performance, and production. *Annual Review of Applied Linguistics*, *36*, 136–163. https://doi.org/10.1017/S0267190516000039

Appendix A. Topics, Communicative Functions/Tasks, and Key Grammar of the Course Curriculum

Communicative Functions/Tasks & Key Grammar	**Use of *Iruda*** (Yes / No)
Topic 1. Birthday party & small events	
• Planning events	
• Talking about past experiences	
• Communicating in Intimate Style	
• 반말 (*banmal*, half-talk; Intimate Style)	Yes
• Verb connective particle for causation or reasons: – (으)니까 (because)	No
• Verb nominalization: – (으)ㄴ/는 것 (the fact that one does/is)	Yes
• Prohibitive or negative imperative verb form: – (으)면 안 되다 (must/should not)	No
• Adverbial clauses/phrases for concurrent or continuous action: – 는 동안(While)	No
• Past tense for past experiences: – (으)ㄴ 적이 있다/없다 (have/have not past particles)	Yes
Topic 2. Korean food	
• Ordering and recommending food	
• Describing food flavors	
• Talking about recipe	
• Future-tense verb conjugation for intention: – (으)ㄹ래요 (will)	No
• Verb conjugation for causation or reasons: – 거든요 (That's because)	No
• Verb conjugation as a modifier of an evaluative noun: – (으)ㄹ 만하다 (Be worthy of doing)	Yes
• Verb conjugation/verb phrase for future reference or use: – 어/아 놓다 (Do for future reference or use)	No
• Noun-modifying form (future action): – (으)ㄹ (to verb)	No
•Verb conjugation/verb phrase for suppositions: – (으)ㄴ/는/(으)ㄹ 것 같다 (it seems)	No
Topic 3. College life	
• Talking about college life	
• Talking about plan and decision	
• Talking with close friends	
• 반말 (*banmal*, half-talk; plain style)	Yes
• Verb conjugation/verb phrase for tendency: – (으)ㄴ/는 편이다 (tend to)	Yes
• Noun phrases with interrogatives: 어떤/무슨 – (이)든지 (any/whichever/whatever [noun])	No
• Verb nominalization/verb phrases for decision/plan: – 기로 하다 (decide to)	Yes
• Noun connective for temporal order: [Noun] – 전에 (before, ago)	Yes
• Adverbial suffix: – 게 (to verb)	Yes

Topic 4. Leisure activities • Talking about leisure activities • Expressing tentative plan • Expressing wish	
• Verb conjugation/verb phrase for appearance or suppositions: – 어/아 보이다 (it looks)	Yes
• Verb conjugation/verb phrase for tentative plans: – (으)ㄹ까 하다 (might/may do)	No
• Verb conjugation/verb phrase for wishes and hopes: – 았/었으면 좋겠다 (I wish)	Yes
• Verb conjugation for reminding of relevant background: – 잖아요 (you know that)	No
• Particles/topic markers for expectancy: – (이)나 (as much/many as); – 밖에 (only, nothing but, no more than)	No
• Verb connective conjugation for concurrent or continuous action: – (으)면서 (while)	Yes

Appendix B. Example Activity Sheet

Week 7 In-Class Activity. *Whose Summer Plans Are Most Fun?*

Key expressions for the task

• 반말 (*banmal*, half-talk; plain style)

• Verb nominalization/verb phrases for decision/plan: – 기로 하다 (decide to)

• Noun connective for temporal order: [Noun] – 하기 전에 (before, ago)

• Adverbial suffix: – 게 (to verb)

Let's make some fun summer plans with *Iruda*, share them, and vote for the most fun summer plans.

Use the three expressions '– 기로 하다,' '– 하기 전에,' and '– 하고 나서' during your conversation with *Iruda*. Also, make a note of three (or more) new expressions you learn from *Iruda* on each step.

Step 1. Practice '-기로 하다'

Find out what *Iruda* is going to do for the given days. First, talk about your plans and then ask about *Iruda*'s. Based on your conversation, complete the table. Make sure to use the expression '– 기로 하다' in an informal style (반말) as in the example below:

Example A:

You: 나는 이번 주말에 영화 보기로 했어. 이번 주말에 무슨 계획 있니?

Iruda: 나는 테니스 치러 가기로 했어.

When	Your plans ('– 기로 했어')	*Iruda*'s plans ('– 기로 했어')	New expressions?
내일			
모레			
이번 주말			
다음 주말			

Step 2. Make a fun summer plan with *Iruda*

Talk about your plans for this summer with *Iruda*, using the expressions '– 기로 하다,' '– 기 전에(는),' and '– 하고 나서.' Then, compare your plans with *Iruda*'s to finalize a fun summer plan. Post your finalized plan on our *Padlet* page.

Example B

You: 이번 여름 방학에 무슨 계획 있니?

Iruda: 교환 학생으로 한국에 가**기로 했어**.

You: 그렇구나. 한국 가**기 전에는** 무슨 계획 있니?

Iruda: 한국 가**기 전에는** 한국 노래 배**우기로 했어**.

When	Your initial plans	*Iruda*'s plans	Your final plan	New expressions?
여름방학				

Step 3. Vote for the most fun summer plans!

Appendix C. Survey Items

A. Selected-response questions and follow-up open-ended questions

1. On a scale of one to ten, how easy was it to navigate and use *Iruda* as a language learning tool?
2. On a scale of one to ten, how much has practicing with *Iruda* helped you to improve your Korean vocabulary?
3. On a scale of one to ten, how much has practicing with *Iruda* helped you to expand your Korean grammar?
4. On a scale of one to ten, how beneficial do you think practicing with *Iruda* was for reinforcing what you learned in your Korean classes?

4a. Would you elaborate how and why (or why not)?

5. On a scale of one to ten, how helpful do you think it will be to continue to use *Iruda* in future Korean language classes?
6. On a scale of one to ten, how much would you like to recommend *Iruda* to someone who is learning Korean?

6a. Would you elaborate why or why not?

B. Open-ended questions

7. What is the most enjoyable aspect of chatting with *Iruda*?
8. What is the least enjoyable aspect of chatting with *Iruda*?
9. Which areas of Korean language do you think are hard to improve by using or chatting with *Iruda* and why?
10. Any comments on using *Iruda*?

Appendix D. Descriptive Statistics of Students' Responses to Numerical Survey Items

Questions	Mean	Median	Standard Deviation	95% Confidence Interval	
				Lower	Upper
Q1. Easiness to navigate	3.85	4.0	0.77	3.64	4.06
Q2. Usefulness (vocabulary)	3.75	4.0	1.0	3.48	4.02
Q3. Usefulness (grammar)	3.43	3.0	0.93	3.18	3.68
Q4. Usefulness (practicing course content)	4.04	4.0	0.90	3.80	4.28
Q5. Recommended for future classes	3.81	4.0	1.0	3.54	4.08
Q6. Recommended for others	3.74	4.0	0.92	3.49	3.99

Appendix E. A Summary of Responses to Open-Ended Items (Questions, 4a, 6a, and 7 to 10)

Perceived benefits and enjoyable aspects

- A variety of conversation topics for speaking
- Opportunities to use Korean learned in class
- Opportunities to learn new expressions
- Good practice for *banmal* conversations
- Plenty of opportunities for informal chatting with "a local Korean"
- Authentic conversations and settings
- Provision of input through interaction without limit
- Practicing Korean anytime without scheduling
- Interesting Konglish expressions
- Exposure to some dialects and current events
- Practicing typing Korean
- *Iruda*'s fun and approachable personality

Limitations or concerns

- Initial difficulty of understanding instructions in Korean
- Locally used complicated slangs not in dictionary or Google translation
- *Iruda*'s assumption that users are in Korea
- Possibility of misrepresentation of Korean culture
- Too much input (while I can only respond slowly)
- No correction of my Korean

Desired features

- Search options to find previous conversations
- A similar application targeting learners of Korean (rather than native Korean speakers)

About the Authors

Ji-young Shin is Assistant Professor of Education Studies and Language Studies at the University of Toronto Mississauga. Her research focuses on language assessment, CALL/educational technology, corpus linguistics, and quantitative research methodology, appearing in journals such as *Language Testing*, *Language Assessment Quarterly*, and *British Journal of Education Studies*.

E-mail: jiyoung.shin@utoronto.ca

ORCiD: https://orcid.org/0000-0002-8710-5993

Yujeong Choi is an Assistant Professor in the Department of East Asian Studies at the University of Toronto. Her research interests include second language acquisition, Korean language pedagogy, project-based learning, and the integration of educational technology in language education.

E-mail: yujeong.choi@utoronto.ca

ORCiD: https://orcid.org/0009-0008-7713-7482