TYPE Original Research
PUBLISHED 24 February 2026
DOI 10.3389/feduc.2026.1703664

Check for updates

OPEN ACCESS

EDITED BY
Sergio Ruiz-Viruel,
University of Malaga, Spain

REVIEWED BY
Shunan Zhang,
Sungkyunkwan University, Republic of Korea
Wen-Chi Hu,
Chaoyang University of Technology, Taiwan

*CORRESPONDENCE
Alireza M. Kamelabad
✉ alimk@kth.se

RECEIVED 11 September 2025
REVISED 25 December 2025
ACCEPTED 21 January 2026
PUBLISHED 24 February 2026

CITATION
M. Kamelabad A, Turano B, Lundin M and Skantze G (2026) Personalized language learning with an LLM chatbot: effects of immediate vs. delayed corrective feedback. *Front. Educ.* 11:1703664.
doi: 10.3389/feduc.2026.1703664

Personalized language learning with an LLM chatbot: effects of immediate vs. delayed corrective feedback

Alireza M. Kamelabad [1]*, Beatrice Turano [2], Mattias Lundin[1] and Gabriel Skantze [1]

[1]Division of Speech, Music, and Hearing, KTH Royal Institute of Technology, Stockholm, Sweden,
[2]Center for Mind/Brain Sciences (CIMeC), University of Trento, Rovereto, Italy

The emergence of Large Language Models (LLMs) has opened new possibilities for language learning through conversational interaction with chatbots. Yet, little empirical evidence exists on how students experience such interactions and how corrective feedback should be provided. Research suggests that immediate corrective feedback is generally more effective than delayed feedback. Nevertheless, learners' perception of this effectiveness and their preferences for feedback timing, particularly in the domain of Computer-Assisted Language Learning (CALL), remain underexplored. This study investigates the feasibility of providing immediate feedback and examines the impact of feedback timing on user experience and grammar learning gains in English. An in-the-wild experiment was conducted with 66 L2 English learners, who integrated chatbot sessions into their English course as an extracurricular activity over one semester. Participants were randomly assigned to two groups receiving feedback either during or after the conversation. Findings reveal no significant difference in learning gains, but immediate feedback enhanced user experience, leading to overall positive perceptions of the chatbot. Additionally, we explore users' perceptions of the chatbot's social role and personality, offering a roadmap for future enhancements. These results provide valuable insights into the potential of LLMs and chatbots for language learning.

KEYWORDS

chatbot, corrective feedback timing, GPT, large language model (LLM), second language learning, Artificial Intelligence (AI)

1 Introduction

Globalization has increased the demand for learning new languages. Technological advancements have led to the development of assistive tools for language learning, but most of them have been limited in their capabilities. The advent of Large Language Models (LLMs), with their remarkable abilities in understanding and generating natural language, presents unprecedented opportunities for personalized language education at scale. However, it is important to assess these models' potential positive and negative impacts before widespread adoption and to guide the design of more appropriate language learning tools using LLMs in the future.

However, access to such meaningful interaction is often constrained in practice–for many learners, fluent speakers are not readily available, and even when they are, affective factors such as anxiety and fear of negative evaluation can suppress participation, both in conversations with Native Speakers (NSs) and in classroom settings (Macintyre, 2007; Horwitz et al., 1986). In larger classes, students primarily practice with peers, which limits the introduction of new structures without teacher mediation; use of the native language may further reduce Second Language (L2) production, and individualized, timely feedback is scarce (Haines, 1995; Bibauw et al., 2019). These access and affective barriers motivate tools that can provide low-pressure conversational practice with individualized Corrective Feedback (CF) outside class.

To address these constraints, learners increasingly supplement classroom work with mobile and web applications alongside traditional tools (e.g., textbooks, flashcards).The language classroom is not the only way learners can practice their language skills. Besides traditional methods, like textbooks or flashcards, there are mobile and web applications designed for language learning. However, in most cases, these applications do not allow learners to practice those conversational skills that interactionists advocate for (Li et al., 2022). A potential tool to alleviate these problems is chatbots, which are designed to interact and converse with humans at any convenient time. Extensive research is being carried out to understand the use of chatbots as language learning tools (Huang et al., 2022; Fryer et al., 2020). Results show that learning through chatbots is not only possible (Kim, 2018b; Jia et al., 2012), but chatbots also allow the student to overcome some of the issues of talking to a NS or speaking in the classroom. They are usually always available, affordable, and they can provide personalized feedback. The Edubot chatbot, for example, converses with the learner about a chosen subject and gives grammatical feedback on the learner's production at the end of the conversation (Li et al., 2022). With the integration of an LLM, a language learning chatbot could not only engage learners in conversations on a wide range of topics but also provide feedback on their mistakes and potentially adapt to users' individual characteristics, such as preferred type of CF, proficiency level, and timing for CF.

Previous research has documented differences in the timing of CF provision to language learners, impacting both learning outcomes and learner preferences. Most of these studies conclude the superiority of Immediate Corrective Feedback (ICF) to Delayed Corrective Feedback (DCF) (Martinussen et al., 2005; Opitz et al., 2011; Fu and Li, 2019, 2022). However, there are also studies showing that DCF has more positive effects on language learning (Metcalfe et al., 2009). Some of the studies could not find any differences when comparing ICF and DCF (Li et al., 2016). These mixed results do not explain the differences in the timing of CF, and since the previous work on chatbots for language learning has mostly relied on a delayed feedback, the effects of an immediate feedback are yet to be studied.

On the other hand, recent advancements in Natural Language Processing (NLP) and Artificial Intelligence (AI) have led to the introduction of LLMs. These models have proven to be highly effective on a number of NLP tasks, like classification, translation, or summarization (Brown et al., 2020), and they are able to discuss diverse scenarios. A chatbot for language learning equipped with a LLM could therefore not only be able to engage the learner in conversation about a broad range of topics, but also be able to give feedback on the learner's mistakes and potentially adapt to some of the user's characteristics, such as proficiency level. These new technologies enable us to provide ICF automatically and effectively in real time, which was previously impossible due to technical limitations. Chatbots, as one of the simplest forms of conversational AI systems, can be used for this purpose. However, the impact of the timing of CF on learners when delivered by chatbots is unclear and yet to be studied. Crucially, prior research on CF timing has focused exclusively on human tutors, yet the psychological dynamics of human-AI interaction may differ fundamentally. Recent large-scale evidence demonstrates that learners perceive and trust feedback differently based on its attributed source, even when the quality and helpfulness are comparable (Henderson et al., 2025). The mere awareness that feedback originates from AI rather than a human instructor triggers distinct cognitive and affective responses–including differences in perceived trustworthiness, interpersonal risk, and emotional reactions–that influence how learners engage with and process the feedback (Henderson et al., 2025; Jacobsen et al., 2025). Moreover, source-related biases and expertise heuristics can moderate these effects, with learners' perceptions varying systematically depending on whether they attribute feedback to AI, expert, or peer sources (Jacobsen et al., 2025). Investigating these source attribution effects is therefore essential for informing the design of AI-based educational tools and their optimal integration into language learning contexts. This knowledge is essential for designing effective educational systems that use chatbots.

As language learning practices evolve with technology, understanding how learners perceive feedback from LLM-powered chatbots becomes crucial. Studies have shown that the personality of chatbots can significantly impact user engagement and satisfaction (Ruane et al., 2021; Mehra, 2021). This interaction suggests that the perception of a chatbot's personality could play a significant role in the uptake of CF and its timing, influencing language learning outcomes. Therefore, examining these perceptions aligns with the broader aim of optimizing feedback mechanisms in L2 practice through chatbots.

Abbreviations: ADHD, attention-deficit/hyperactivity disorder; AI, Artificial Intelligence; CALL, Computer-Assisted Language Learning; CF, Corrective Feedback; DCF, Delayed Corrective Feedback; EEG, Electroencephalogram; EFL, English as a Foreign Language; GEC, Grammatical Error Correction; GED, Grammatical Error Detection; GLMM, Generalized Linear Mixed Model; ICF, Immediate Corrective Feedback; L1, First Language; L2, Second Language; LLM, Large Language Model; NS, Native Speaker; NLG, Natural Language Generation; NLP, Natural Language Processing; NLU, Natural Language Understanding; SLA, Second Language Acquisition.

1.1 Research questions

In this study, we present an in-the-wild experiment where 66 students interacted with an LLM-powered chatbot over 12 sessions distributed over a one-month period. A between-subject design was used, where we investigated the effects of such a system on users' learning outcomes and perception about the chatbot, as well as a comparison of ICF vs. DCF given by the chatbot. More specifically, our research questions are:

RQ1 How can chatbots using LLMs be utilized for language learning?

RQ2 How is a LLM-powered chatbot for L2 practice perceived by language learners?

RQ3 How does the timing of CF (immediate vs. delayed) impact the user's preference in using LLM-powered chatbots for language learning practice?

RQ4 How does the timing of CF (immediate vs. delayed) impact the grammar learning outcomes?

2 Background and related works

The intersection of language learning and technology has increasingly become an avenue for innovative educational methodologies. This literature review focuses on the crucial role that conversation and interaction play in the acquisition of a L2, particularly in conjunction with technological advancements such as chatbots. CF, a pivotal element of language pedagogy, is examined in depth, with specific attention given to its timing, modalities, and its reception in the context of interactive tools. By synthesizing key findings from previous studies, this review aims to elucidate the significance of these elements in enhancing the efficacy of L2 learning and to establish a framework for the present study.

2.1 Empirical findings on interaction in the Second Language (L2) learning

Leveraging conversational interactions in L2 learning is a dynamic way to facilitate various critical processes that foster language acquisition. Pica (1996) underscored the importance of feedback, modified output, and other cognitive and social operations such as negotiation, collaborative dialogue, and instructional intervention in this context. These processes not only provide essential input and feedback but also create opportunities for learners to produce and refine their L2 knowledge. While interactions with NSs can offer rich, varied linguistic data, challenges may arise, such as inhibiting learner output. However, interactions with fellow learners, despite potential limitations in linguistic resources, can also be beneficial by providing interlanguage data and scaffolding.

The value of conversation in L2 learning extends to interactional feedback, which facilitates the negotiation of meaning and provides comprehensible input (Chen, 2017). Moreover, conversational exchanges promote collaborative dialogue, which is instrumental in driving language proficiency. Complementing this, Ammar and Hassan (2018) found that conversations focusing on linguistic forms provide significant opportunities for co-constructing grammatical knowledge–effects that are particularly pronounced for learners with lower proficiency, with variations dependent on task type and linguistic focus.

Nevertheless, the path to harnessing the full potential of conversational interaction is not without obstacles. Insufficient focus on form, which affects learners' attention to the accuracy during communication tasks and may result in inadequate feedback, are among the shortcomings that are identified (Philp et al., 2010). Social dynamics also come into play, where proficiency levels, personality, peer relationships, and task orientation can all influence one's engagement in language-related episodes and the effectiveness of CF during peer interactions. Furthermore, the limitations inherent to interacting with less proficient peers can lead to missed opportunities for negotiation due to a lack of linguistic resources. Endorsing the role of NSs in this scenario, Dörnyei (2007) emphasized their contributions in providing critical feedback and corrections, aiding learners in circumventing language transfer errors.

Additionally, teacher-student interactions characterized by 'Teacher Talk' have been found to be particularly effective when they encompass informal dialogues, fostering a deeper understanding of English grammar (Alkhazraji, 2018). Practical speaking exercises are also instrumental in improving students' oral grammatical production (Wahyuni and Afrianti, 2021). These interactions expose learners to a diverse array of L2 linguistic registers, thereby nurturing their communicative competence—the ability to use language appropriately in context (Hymes, 1972)—as further detailed in Section 2.2 communicative competence (Yorozu, 2001). The explicit clarification of grammar rules can be beneficial, especially for adult learners, although the pairing of such explanations with ample practice is essential to attaining fluency (Campbell, 1970). NSs naturally provide well-formed input through conversational adjustments that accommodate non-native speakers, thereby facilitating the acquisition of language patterns (Long, 1983). Saito's study suggested that regular dialogue with NSs could lead to improvements in learners' linguistic capabilities across various domain, including grammar (Saito and Akiyama, 2017). Extending beyond informal interaction, "processing instruction" by fluent speakers–activities that prompt learners to actively relate grammatical form to meaning–has been recognized as a powerful method to grow grammatical systems in L2 learners (Jelinski and VanPatten, 1997).

The synthesis of these studies illustrates that conversational interaction is a multifaceted tool in L2 learning, offering numerous benefits while also presenting challenges that must be navigated with thoughtful consideration of factors such as the type/timing of feedback provided and the nature of the interaction. The following section introduces the theoretical frameworks that explain why these interactional processes facilitate language acquisition.

2.2 Theoretical frameworks

Theories linking interaction with L2 learning illuminate the essential mechanisms underlying successful language learning. Central to these is Interaction Hypothesis, which posits that language acquisition is enhanced through interactive communication and the negotiation of meaning (Long, 1981). Long (1981) argues that conversational engagement where learners and their interlocutors collaborate to understand and convey messages allows learners to interact within their Zone of Proximal Development, benefiting from input that facilitates linguistic development. Other significant theories complement Long (1981)'s findings, such as the "Comprehensible Output" hypothesis, which underscores the importance of language production in communication for language learning (Swain, 2005; Swain and Suzuki, 2008). Swain (2005)'s emphasis on language output reflects the necessity for learners to actively use the language, thereby solidifying their understanding of its structures.

The "Sociocultural" theory places interaction within the sociocognitive domain, positing that learners progress in their linguistic abilities through dialogue with more proficient speakers (Vygotsky, 1978). This process, which occurs within the learner's Zone of Proximal Development, indicates that conversational interaction is not merely facilitative but intrinsically tied to cognitive development. Additionally, Hymes (1972) introduced the idea of "Communicative Competence", which emphasizes the practical use of language in context, beyond grammatical correctness, a skill honed through authentic interactional experiences (Hymes, 1972). These theories converge on the premise that the intricate interplay between conversational interaction, input, output, and the salience of linguistic features coalesce to form the bedrock of L2 learning and acquisition.

2.2.1 Noticing hypothesis

On top of all, the "Noticing" hypothesis also supports the value of interaction by positing that conscious awareness of linguistic features is a prerequisite for their acquisition (Schmidt, 1990, 2001). The Noticing hypothesis posits that the act of consciously noticing linguistic features within input is critical for learners to acquire those features. The hypothesis suggests that learners must first be aware of the forms and structures of the language before they can process and integrate them into their own interlanguage system. In the realm of conversation and interaction, the Noticing Hypothesis is particularly relevant. Engaging in dialogues with interlocutors, especially those that involve negotiation of meaning or CF, brings learners' attention to specific linguistic forms (Swain, 1985; Long, 1981). This may happen when a learner encounters a new word, structure, or when their output is corrected, prompting them to notice the gap between their current interlanguage and the target language norm.

Schmidt (1990)'s theory aligns well with aspects of interactional feedback, which can aid learners in noticing errors in their language production (Long, 1981). For example, during a conversation, a learner might not realize they are using an incorrect tense until the interlocutor either *implicitly* or *explicitly* points it out. The moment of realization–the noticing–can then lead to language development as the learner attempts to adjust their output in subsequent interactions. Furthermore, Schmidt (1990) contended that attention is a key cognitive process in learning and that noticing is a byproduct of attention (Schmidt, 1990, 2001). Thus, the interactions that draw learners' attention to language form facilitate the process of noticing. It is through this process that interaction becomes not just a medium of communication but a dynamic environment for focused language learning and acquisition.

2.3 Corrective feedback and implications of noticing hypothesis

CF in Second Language Acquisition (SLA) refers to the responses provided to a learner's production that deviates from the correct forms in the L2 (Li, 2010). It offers *negative evidence*, which contrasts with *positive evidence* that consists of correctly formed structures in the L2. Some theories contend that positive evidence should be sufficient for SLA (Krashen, 1981; Truscott, 2007). However, interactionist perspectives, as advocated by Long and Pica, contend that negative evidence and subsequent corrective actions play a crucial role in learning as they guide learners to notice and address their linguistic errors (Long, 1996; Pica, 1988).

Havranek (2002)'s findings further emphasize that CF not only aids in the rectification of erroneous structures but also fosters a heightened correct application of those structures in future language use, beyond the immediate feedback context (Havranek, 2002). Additionally, CF acts as a catalyst in heightening learner's conscious awareness of language norms, thus enabling them to actively identify their linguistic shortcomings and engage in autonomous self-correction (Penning de Vries et al., 2011). This emphasizes a more learner-centred approach to CF, where the position of the learner involves internalizing feedback and integrating it into their evolving linguistic framework. Further, feedback provided by a NS or computers was found to be more effective than feedback provided by language teachers (Li, 2010).

2.3.1 Typology of corrective feedback, their relative effectiveness, and personalization

CFs are often divided into primarily two encompassing categories: *Explicit* and *Implicit* CF (Ellis et al., 2006; Ellis, 2011, 2021; Li, 2023).

*Implicit CF*subtly indicates errors without explicitly marking them, encouraging learners to self-identify and correct their mistakes. This category often includes *recasts*, which are reformulations of the learner's utterance without overtly indicating errors. However, they might not always lead the learner to notice mistakes due to the absence of explicit correction cues (Ellis et al., 2006). Other Implicit forms include *repetition*, where the error is repeated to draw attention; *clarification requests*, prompting the learner to reconsider their utterance; *elicitation*, encouraging the learner to correct the error themselves; *paralinguistic signals*, non-verbal cues hinting at mistakes; *metalinguistic interpretation*, providing hints or questions about the error; and *interruption*, halting the learner to signal an error (Li, 2023).

*Explicit CF*Overtly identifies the error, offering direct routes for correction. It encompasses *meta-linguistic feedback*, where information is provided on the incorrect grammar rule or linguistic structure used by the learner; and *explicit correction* in oral or conversational contexts, overtly indicating the mistake to the learner. In written contexts, *direct feedback* supplies the correct form, while *indirect feedback* indicates the error location without specifying the correction, and *meta-linguistic feedback* points out the error type or offers a rule without providing the correct form (Ellis, 2009, 2021).

Additionally, CF strategies can be oriented toward requiring learners to either *revise* the text, incorporating the given feedback, or merely draw *attention to correction*, acknowledging the feedback without necessitating text revision. This strategic distinction aims to facilitate learners' engagement with feedback in a manner aligned with instructional goals and individual learning processes.

The effectiveness of oral CF mechanisms in L2 learning has been a focal point of empirical research, revealing nuanced outcomes based on the type of feedback and the context of its application. Studies indicate that explicit forms of oral CF, such as explicit correction and metalinguistic feedback, substantially enhance learners' ability to recognize and rectify errors in their utterances. This is attributed to the overt nature of the feedback, which facilitates learners' noticing of errors, an essential step for language acquisition (Rassaei, 2013; Li, 2010; Norris and Ortega, 2000). In contrast, implicit feedback types, including recasts and elicitation, have demonstrated varied effectiveness, largely hinging on learners' ability to discern the corrective intent behind the teacher's input. Recasts, despite their widespread use, often fall short in promoting learner noticing due to their subtlety, limiting their immediate effectiveness compared to more explicit interventions (Ellis et al., 2006; Sheen, 2010). However, certain studies, such as those by Li, have observed a slightly larger long-term effect for implicit feedback, suggesting its potential in fostering deeper linguistic processing over time (Li, 2010). Interestingly, prompt-based strategies, including elicitation and clarification requests, have shown efficacy in eliciting lexical repairs, indicating their value in encouraging active learner engagement with the feedback process (Tan et al., 2022).

Similarly, the landscape of written CF presents a rich tapestry of strategies whose effectiveness is closely linked with their explicitness and focus. Direct feedback, characterized by directly providing the correct form, alongside indirect feedback, which signals the presence of an error without specifying the correction, have both shown promise in improving learners' written accuracy in the short term (Beuningen et al., 2008; Norris and Ortega, 2000; Brown et al., 2023). Among these, metalinguistic feedback, providing learners with hints or rules about their errors, has been singled out as particularly effective, especially when coupled with metalinguistic comments, underscoring the significance of engaging the learner's cognitive processes in the correction journey (Ellis, 2009; Brown et al., 2023). This aligns with Bitchener's findings, which suggest that targeted, explicit CF on specific linguistic features can lead to significant, lasting improvements in accuracy, underlining the efficacy of focused feedback approaches for lasting language development (Bitchener, 2008). The overarching consensus from recent meta-analyses and literature reviews reinforces the superiority of feedback types that are explicit and targeted, as these are more readily noticed, comprehended, and integrated by learners, thereby accelerating the L2 acquisition process (Li, 2023; Norris and Ortega, 2000).

2.3.2 Importance of personalization and adaptation in corrective feedback

The vital role of personalization in CF becomes increasingly evident when considering the distinct requirements of individual learners in communicatively oriented L2 classrooms. While studies have shown that techniques like elicitation and metalinguistic feedback, compared to more common practices such as recasts, result in higher rates of learner uptake and negotiation of form, it is the customization of these feedback mechanisms to align with the learners' proficiency levels that significantly enhances the effectiveness of CF (Lyster and Ranta, 1997). Moreover, affective factors such as learners' orientation, attitudes, beliefs, goals, and strategies play a crucial role in how feedback is processed and retained, emphasizing the necessity to adapt CF strategies to individual motivational and psychological profiles (Storch and Wigglesworth, 2010).

Optimal language learning outcomes are further facilitated through the careful adaptation of CF to learners' unique characteristics, including motivation, personality, aptitude, learning style, and preferred language learning strategies (Penning de Vries et al., 2011). Acknowledging that different learners may favor various types and modes of CF, and respond distinctly to these interventions, underscores the importance of tailoring CF not only to enhance noticing and conscious awareness of language features but also to accommodate the developmental readiness and learning styles of individual learners (Penning de Vries et al., 2011; Havranek, 2002). Thus, personalizing CF not only cultivates optimal learning conditions but also augments learner satisfaction and motivation, paving the way for more effective grammar acquisition and learner engagement in the process of L2 acquisition (Penning de Vries et al., 2011).

2.3.3 Timing of corrective feedback

CF in SLA can be temporally classified into *ICF* and *DCF*, delineating when feedback is provided in relation to learner errors. ICF is administered directly following an error, either in oral or written communication. Conversely, DCF is delivered after a lapse of time, which could span from the end of a task to subsequent class sessions or even days, introducing periods of delay between the error occurrence and feedback provision (Quinn, 2021; Fu and Li, 2022). In the context of communicative tasks in SLA, the immediate and delayed feedback are also referred to as Online and Offline feedback, respectively; where the CF either happens right after a mistake is made during the communicative task and includes interruption in the flow of the conversation and the offline feedback includes the feedback after the communicative task has finished. In this paper we use Online and Offline feedback interchangeably with ICF and DCF (Fu and Li, 2022). Different theoretical perspectives can account for how immediate and delayed CF facilitate L2 development. Quinn (2021) in details review the relevance of these theoretical frameworks to the CF timing: Sociocultural theory and skill acquisition theory suggest that ICF can be more effective as it provides direct social interaction and opportunities for practice. On the other hand, theories of cognitive comparison and reactivation and reconsolidation propose that both immediate and delayed CF can be beneficial as they allow for memory encoding, retrieval, and restructuring.

The empirical research in L2 learning presents a diverse landscape regarding the effectiveness of ICF and DCF, with a general trend favoring the former. Studies consistently showcase the superiority of ICF in nurturing language development, attributing this advantage to its alignment with communicative immediacy and the interaction hypothesis (Li, 2010; Long, 1996). Immediate feedback facilitates the acquisition of implicit knowledge and promotes cognitive engagement through direct error correction and practice (Opitz et al., 2011; Fu and Li, 2022). Moreover, the synchronous nature of ICF, especially in written communication, has been linked to improved accuracy and error reduction, highlighting its role in engaging cognitive and social processes conducive to L2 acquisition (Shintani and Aubrey, 2016; Candel et al., 2020).

Conversely, DCF's effectiveness is nuanced, dependent on factors such as the communicative modality, feedback explicitness, and timing duration (Xu and Zeng, 2023). While some studies suggest that delayed feedback, particularly when contextualized, can rival the effectiveness of ICF (Canals et al., 2021; Quinn, 2021), others argue for its potential cognitive load implications, which could disproportionately affect learners with specific working memory limitations (Fu and Li, 2019) (e.g., attention-deficit/hyperactivity disorder (ADHD)). Nonetheless, the enhanced metacognitive accuracy observed with immediate feedback reinforces its benefit in fostering learner engagement and error processing (Candel et al., 2020). Despite these findings, the construct of 'delayed feedback' warrants further exploration to clarify its parameters and examine its varied impacts comprehensively (Xu and Zeng, 2023).

2.4 Computer-assisted language learning (CALL)

Computer-Assisted Language Learning (CALL) has been proven to provide numerous benefits for L2 learners (Dina and Ciornei, 2013). In a meta-analysis, it was found that computer/technology-supported language learning is at least as effective as instruction without technology, and in studies using rigorous research designs, the CALL groups outperformed the non-CALL groups (Grgurovic et al., 2013). These findings suggest that computers and modern tools benefit language learning. One particular subcategory of CALL is *dialogue-based* CALL systems (Bibauw et al., 2019). These systems utilize dialogue between the learner and the system as the primary means of learning. This approach assumes that computers can provide meaningful conversational practice in the target language, which, according to the interactionist approach to SLA, is essential for language learning (Long, 1996). Examples of these systems include social robots, conversational agents, and chatbots specifically designed or adapted for this purpose. The explicit and prominent nature of CF, particularly when providing an explanation and highlighting errors in written grammar and vocabulary exercises within a CALL environment, significantly influences learner uptake, with Metalinguistic plus Highlighting feedback type showing the highest uptake and Repetition + Highlighting demonstrating the least (Heift, 2004).

2.4.1 Chatbots for language learning

Chatbots are text-based dialogue systems designed for conversing with humans. One of the earliest chatbots, ELIZA, was introduced as a virtual psychoanalyst (Weizenbaum, 1966). Despite its simplicity, ELIZA's popularity catalyzed further research on chatbots. For many years, chatbots used rule-based architectures to generate linguistic output, resulting in repetitive conversations and limited functionality. With the recent advancements in Large Language Model, more sophisticated chatbots have been developed. A language model assigns probabilities to word sequences (Mikolov et al., 2013). Advancements in computational power and the availability of large text corpora have led to the development of LLMs like GPT-3 (Brown et al., 2020), RoBERTa (Liu et al., 2019), and BERT (Devlin et al., 2019). LLMs capture language patterns at a large scale and generate coherent text with human-like fluency (Radford et al., 2019; Brown et al., 2020), which was not the case with simpler language models. They handle complex tasks such as text summarization/generation, and question answering (Radford et al., 2019). LLMs comprehend language and context, making them suitable for various NLP applications (Radford et al., 2019) such as Natural Language Understanding (NLU), Natural Language Generation (NLG), Grammatical Error Detection (GED), and Grammatical Error Correction (GEC). In chatbots, LLMs provides a natural and human-like conversational experience. LLMs excel in *Context Awareness*, understanding conversation contexts for relevant responses (Michel et al., 2019). The rise of LLM-powered chatbots, such as ChatGPT (OpenAI, 2022), showcases the ability of these models to generate human-like responses that seamlessly integrate with the conversation history and context. These make them suitable for conversational language learning applications, offering appropriate feedback based on conversation context (Michel et al., 2019).

In the educational domain, chatbots offer several unique advantages compared to traditional methods (Pérez et al., 2020). Firstly, they are always available to meet users' needs, unlike human counterparts or tutors (Huang et al., 2022; Fryer et al., 2020). Additionally, chatbots can be designed to serve as tutors or assist teachers by handling repetitive tasks, relieving the burden on humans (Fryer et al., 2019). They can also serve as innovative learning tools, such as teachable agents (Chase et al., 2009). Furthermore, chatbots provide tailored feedback on learners' performance, which is not always available in traditional classrooms (Ruan et al., 2021).

In the field of L2 learning, chatbots have gained increasing popularity over the last decade, with many studies highlighting their advantages (Haristiani and Rifa'i, 2020; Pham et al., 2018; Kerlyl et al., 2007; Haristiani, 2019; Coniam, 2014; Dokukina and Gumanova, 2020; Li et al., 2022). For example, EduBot is an online language learning tool that engages users in English conversations on various topics and provides delayed grammatical feedback (Li et al., 2022). Another study focuses on improving chatbot conversations with language learners using deep neural networks (Tu, 2020). Additionally, BookBuddy transforms books into engaging conversations using a voice chatbot, resulting in positive engagement among young Chinese learners of English (Ruan et al., 2019). While chatbots show positive effects in certain areas, such as vocabulary acquisition and listening skills, their impact on reading comprehension appears to be minimal (Kim, 2018b,a). Students using chatbots for 1 h per week over a school term performed better in vocabulary acquisition compared to a control group (Jia et al., 2012). Another study showed that students using the chatbot Elbot for 16 weeks outperformed the control group in listening skills but not in reading skills (Kim, 2018b). Furthermore, improvement in grammar was observed in a study using the same chatbot over the same period (Kim et al., 2019).

Many studies highlight the distinct advantages of using chatbots in language learning. They can alleviate the stress, anxiety, and shyness often experienced by L2 students when conversing with fluent speakers or NSs, thus positively impacting language learning outcomes (Ayedoun et al., 2015; Bao, 2019; Ruan et al., 2019; Nakaya and Murota, 2013; Fryer and Carpenter, 2006; Ayedoun et al., 2019). Furthermore, chatbots can provide comprehensive language information that may not be available in peer conversations (Fryer et al., 2019). By continuously tracking users' mistakes and language performance, chatbots can offer real-time tailored feedback to meet individual needs (De Gasperis and Florio, 2012). This adaptability and feedback capability make chatbots versatile language learning tools. Providing CF on the user's desired topic makes chatbots unique compared to most other assistive devices for language learning. There are different ways in which chatbots can provide feedback depending on their design and aim. Edubot, for example, gives two types of feedback: grammatical feedback at the end of the conversation and "pragmatic" feedback during the conversation (Li et al., 2022). However, in the classroom, teachers can give CF both during (ICF) and after (DCF) the interaction with the student.

Different chatbot designs allow for the implementation of different types of feedback. However, there is little experimental research on the effectiveness of these strategies when applied in chatbots. Furthermore, the adequacy of the CF offered lacks a thorough discussion, with scarce specifications regarding the type of grammatical feedback provided by the chatbots.

2.4.2 Recent developments in generative AI for language learning

Since our data collection in early 2022, the release of ChatGPT has accelerated research on generative AI in language education. Recent empirical studies have begun to examine LLM-powered tools for L2 learning, with mixed findings regarding their effectiveness compared to human instruction. Wang (2024) observed that learners receiving AI-generated feedback exhibited lower writing anxiety and greater improvements in complexity and fluency than those receiving teacher feedback, attributing this to the non-judgmental, low-stakes environment that AI provides. However, Soyoof et al. (2025) found that students receiving scaffolded corrective feedback from teachers significantly outperformed those receiving feedback from ChatGPT, with qualitative analyses suggesting that trust in the teacher—driven by personalized emotional and technical support—accounted for this advantage.

In comparing AI-generated to human feedback more broadly, a comprehensive meta-analysis by Kaliisa et al. (2026) synthesizing 41 studies found no statistically significant difference in learning outcomes between the two sources, suggesting functional equivalence in immediate performance. However, learner perceptions reveal a persistent "trust gap." Henderson et al. (2025) report that while students value AI feedback for its accessibility and timeliness, they consistently rate human feedback as more trustworthy. Similarly, Zhang et al. (2025) found that students rated AI feedback highly for usefulness when the source was blinded, but presented a strong bias against it when the AI source was disclosed, highlighting that trust remains a critical psychological barrier despite comparable pedagogical efficacy.

Regarding feedback timing, Qun (2025) examined immediate vs. delayed feedback in online L2 education and found that both significantly enhanced motivation and learning outcomes compared to no feedback, though no substantial differences emerged between immediate and delayed conditions—a finding that aligns with our own results. Cheng and Xu (2025) explored synchronous written corrective feedback in technology-enhanced contexts, finding that learners displayed positive engagement across affective, behavioral, and cognitive dimensions, with factors such as digital literacy and teacher-student relationships mediating engagement. These recent studies provide an emerging empirical foundation for understanding AI-mediated feedback, to which our pre-ChatGPT investigation contributes foundational evidence on feedback timing effects.

2.4.3 Perception of chatbot's personality and corrective feedback's timing

The literature indicates a clear interest in understanding how chatbot personalities influence user interactions across various domains, including language learning (Ruane et al., 2021; Shumanov and Johnson, 2021; Mehra, 2021; Kang, 2018; Zogaj et al., 2023; Nißen et al., 2022; Kuhail et al., 2022). In the present study, we operationalize chatbot personality as the set of anthropomorphic attributes—such as perceived gender, age, and social role—that users spontaneously project onto the conversational agent. This reflects the degree to which the chatbot is experienced as a social actor rather than a mere software tool, consistent with the Computers Are Social Actors (CASA) paradigm (Nass and Moon, 2000). However, there is a noticeable gap in studies directly examining how these perceptions influence the effectiveness and preference of CF timing in language practice with chatbots. The existing research demonstrates users' preferences for chatbot personalities and highlights the potential impact of personalized feedback delivery (Kim et al., 2020; Rassaei, 2023; Wiboolyasar and Jinowat, 2020). Furthermore, it is evident that factors like the timing of feedback and the interplay with chatbot personality could significantly affect language learners' preferences and learning outcomes (Cornillie et al., 2012; Baz et al., 2016). These insights underscore the importance of examining chatbot personalities within the context of L2 learning, particularly how it affects feedback's uptake and timing preference, which remains a critical and underexplored aspect of technology-enhanced language learning.

3 Chatbot architecture

We developed a web-application chatbot utilizing the OpenAI GPT-3 (Brown et al., 2020) API, implemented in `Python 3.8` and `JavaScript`, deployed on our private server via `Python Flask`, and with data stored in a database[1]. The chatbot retrieves initial prompts for each conversational session from the database and initiates the interaction accordingly. We designed our chatbot to simulate written conversation practices between an English instructor (chatbot) and an English as a Foreign Language (EFL) learner (user), where the teacher corrects the user's grammatical mistakes. The CF can occur either immediately after the mistake is made during the conversation or as a final report summarizing the errors and corresponding corrections following the chat. In this section, we will detail the architecture of our language learning chatbot as shown in Figure 1.

3.1 Interaction flow

Users are provided the link to the chatbot system, which they can open in their browser of choice. They then sign up and log in to the experiment dashboard, which displays information regarding the experiment and the available chat sessions. The chat page features text input, a send button, a messages stack, and a completion bar (see Figure 2a as an example of the chat interface). The conversation is stored as an accumulative prompt used to generate context-dependent corrections and the chatbot's GPT-generated responses. The prompt uses two tags to parse the user and agent utterances: `[Student's Name:]` and `[Teacher:]`.

3.1.1 Start of the conversation

The session begins with the chatbot greeting the user and presenting a pre-determined prompt on the conversation topic. The prompt is built upon throughout the session using the user's input and the program's generated responses.

3.1.2 Input processing and output generation

Prior to the experiment we piloted the system to fine-tune the use of GEC and NLG by the LLM. The best result was for utterances no longer than 250 characters. Therefore, for each message the user has a limit of 250 characters. The user input is passed to the GEC component that checks for grammatical mistakes using Gramformer (Prithivida, 2023). To ensure that the correction considers the context of the sentence, a set number of characters of the conversation history is also passed to the GEC. If there are no grammatical mistakes, the system continues to generate a GPT response. Otherwise, it would proceed to `Error Classification` module. Next, the mistake is classified into 16 different grammatical mistake categories (Bryant et al., 2017) that were used by the Gramformer system. In the end, the user error, its classification, and the correction are stored in the database.

We defined three different types of grammatical correction feedback. In case of several mistakes in one utterance, only the first mistake is corrected and shown to the user as a grammatical correction to reduce the cognitive load. The other reason why to choose the first mistake in the utterances was due to the primacy effect. The serial position effect suggests that words or phrases at the beginning (primacy effect) and end (recency effect) of a sequence are better recalled, impacting sentence structure and retention in language processing (Atkinson and Shiffrin, 1968; Baddeley, 2000; Baddeley et al., 2009). This effect shows that items in the middle are often forgotten due to cognitive load, emphasizing the need to structure information for better comprehension (Gernsbacher, 1990). The system randomly chooses what type of grammatical feedback to give the user and generates a CF according to the feedback type and the experimental condition. The feedback types are:

- **Immediate Corrective Feedback (ICF)**: In this condition, participants were corrected immediately after committing mistakes during the conversation. The chatbot subsequently continued the conversation, as illustrated in Figure 2a.
- **Delayed Corrective Feedback (DCF)**: Participants were not corrected during the conversation in this condition. Instead, they received a summary of their errors and corrections immediately following the conversation. This report was displayed only once and was not accessible later. Figure 2b shows an example of a summary report.

The corrections in the DCF are produced with the same algorithm as ICF and saved in the background. At the end of the chat session they are shown to the user as a summary report of their mistakes. Since the same system was used for error detection, the amount of feedback generation was the same between the two conditions, with the only difference being the time when the feedback was given to the participant (immediately in the conversation, delayed in a report). The CFs are shown in random order.

If a new CF was created, it was included in the prompt along with the user's message. Otherwise, only the latest user's message was added to the prompt. The updated prompt is then sent to the GPT-3 API to generate a response. The response is added to the prompt and shown to the user. The conversation ends when the user has sent more than 1000 characters. The character goal

[1] The GitHub repository contains the full implementation code of our project. https://ali.mk/ChatBot2023.

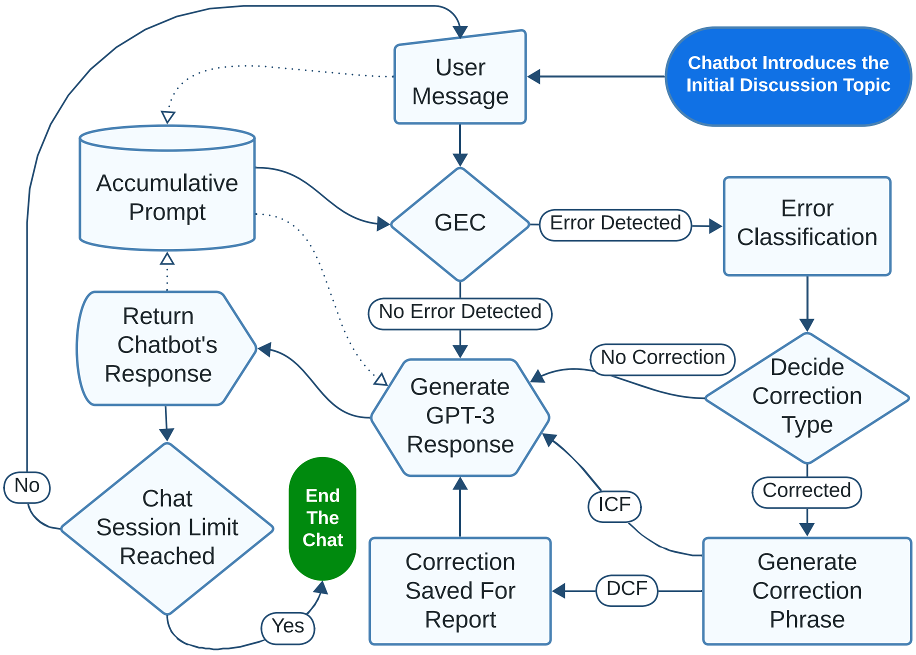


FIGURE 1
Flowchart describing how the agent response is generated from user input.

is weighed to consider factors such as user fatigue, GPT-3 ability, and time limit. It is visually presented to the user as a progress bar. Once the limit has been reached, the user is redirected to the website's homepage.

3.2 Main components

3.2.1 Natural language understanding and generation

The user's input and the chatbot's response are both processed by OpenAI's text-davinci-003 model[2] Brown et al. (2020), which handles the NLU and NLG components. The NLU checks if the input is valid and not empty or composed of special characters. Then, it passes the input to the GED, explained below, to detect and correct grammatical errors. The NLG component uses OpenAI's Completion endpoint to generate a sentence based on the entire conversation history between the user and the bot.

3.2.2 Grammatical error detection and correction

The GEC component utilizes the T5 Tokenizer to process user input and generate a grammatically corrected version using T5 for Conditional Generation (Raffel et al., 2020). The resulting output is then compared to the original input to identify any grammatical errors. In case of a discrepancy, the GED component comes into play, employing Gramformer to locate errors within the sentence (Prithivida, 2023). Gramformer initially detects and classifies the error based on the grammatical error classification proposed by Bryant et al. (2017).

The output of this phase consists of a tuple with seven elements: *CAT*, the category of the mistake (e.g., verb tense, word order, article); *E*, the token identified as the erroneous word; S_e and E_e, the initial and final indices of the error word in the user's sentence; *C*, the token representing the correction of the error word; S_c and E_c, the initial and final indices of the correction in the corrected sentence. The Gramformer model returns a tuple for each error in the sentence. Errors belonging to the categories *PUNCT*, *NOUN*, *ORTH*, and *OTHER* are excluded.

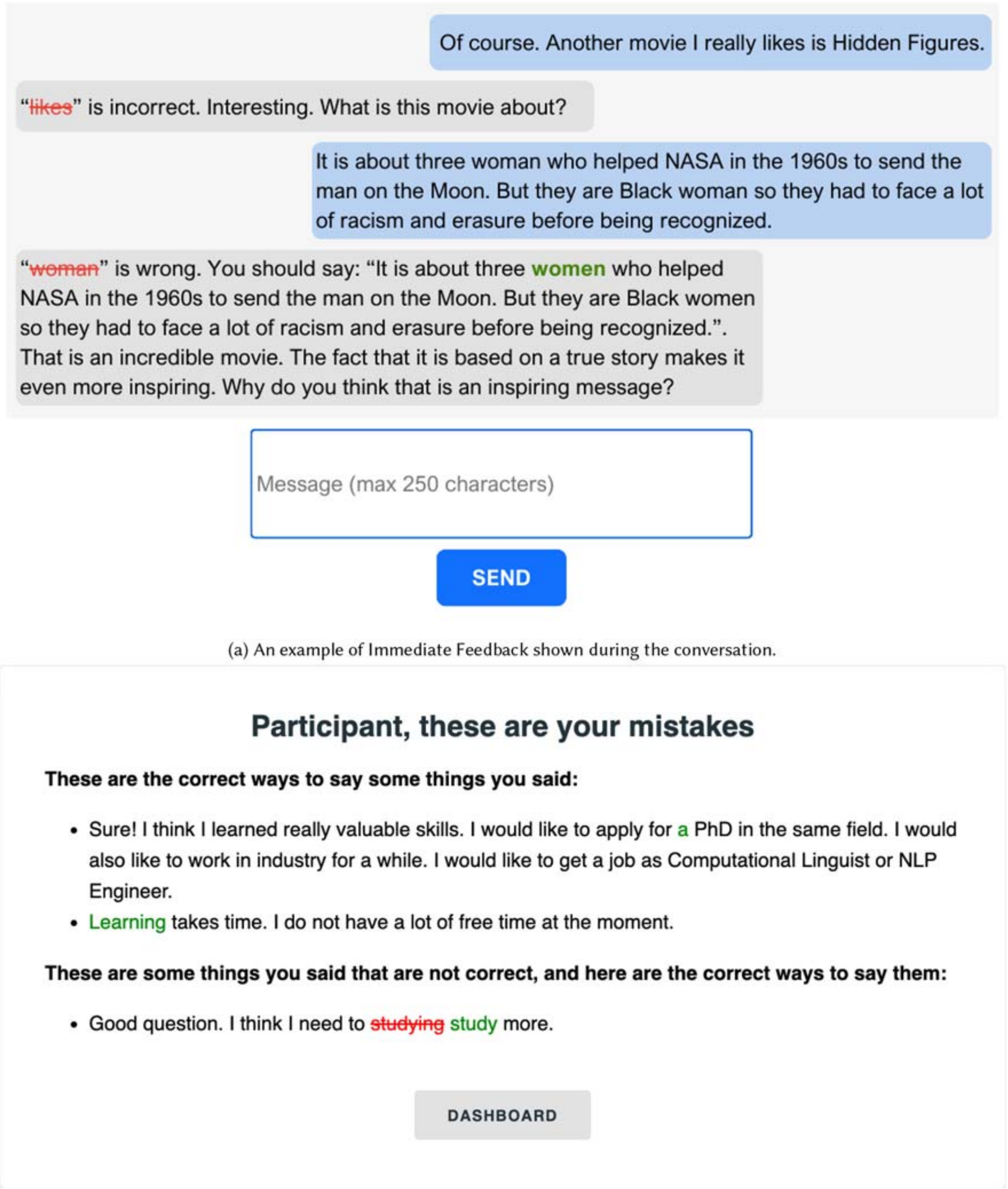
3.2.3 Corrective feedback generation

Upon identifying a grammatical mistake, the bot generates a CF using a custom template tailored to different feedback types:

- **Explicit Recast**. The user receives the grammatically corrected version of their original utterance, marked with green color.
- **Negative Evidence**. The user is provided with information regarding their mistake, specifically pointing out the incorrect word within the sentence marked by the red color.
- **Combined Feedback**. This feedback type merges the previous two, offering the user information about the error and the corrected version of their utterance.

The participants in ICF condition receive the generated feedback right away and for the DCF condition, they are

[2] At the time the study was conducted in late 2022, this was the most advanced model available to us.



Of course. Another movie I really likes is Hidden Figures.

"~~likes~~" is incorrect. Interesting. What is this movie about?

It is about three woman who helped NASA in the 1960s to send the man on the Moon. But they are Black woman so they had to face a lot of racism and erasure before being recognized.

"~~woman~~" is wrong. You should say: "It is about three women who helped NASA in the 1960s to send the man on the Moon. But they are Black women so they had to face a lot of racism and erasure before being recognized.". That is an incredible movie. The fact that it is based on a true story makes it even more inspiring. Why do you think that is an inspiring message?

(a) An example of Immediate Feedback shown during the conversation.

Participant, these are your mistakes

These are the correct ways to say some things you said:

- Sure! I think I learned really valuable skills. I would like to apply for a PhD in the same field. I would also like to work in industry for a while. I would like to get a job as Computational Linguist or NLP Engineer.
- Learning takes time. I do not have a lot of free time at the moment.

These are some things you said that are not correct, and here are the correct ways to say them:

- Good question. I think I need to ~~studying~~ study more.

(b) An example of Delayed Feedback, which is shown after the conversation has ended.

FIGURE 2
(a) An example of Immediate Feedback shown during the conversation. **(b)** An example of Delayed Feedback, which is shown after the conversation has ended. The two types of feedback the system can provide.

accumulated and added to the final report shown after the session ends.

4 Methodology

4.1 Participants

The study enrolled 66 bachelor students from five diverse fields: Computer Science, Economics and Management, Applied Informatics, Informatics and Management, and Information Technology. Of these, 54 participants successfully completed all experiment steps and were included in the analysis. Among the analyzed participants, 24 were assigned to the Immediate Corrective Feedback (ICF) condition and 30 to the Delayed Corrective Feedback (DCF) condition; this imbalance resulted from differential attrition (see Section 5.5).

The analyzed sample comprised 43 males and 11 females, with ages ranging from 18 to 23 years ($\bar{x} = 20.59$, $s = 0.98$). All

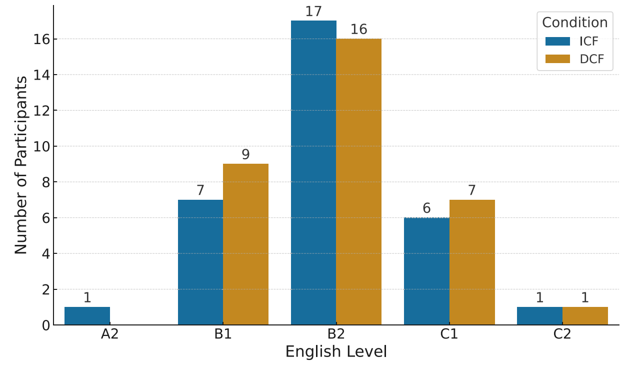


FIGURE 3
Number of participants by English level in different conditions.

participants were native Czech speakers, apart from one Chinese speaker.

Prior to the experiment, participants self-assessed their English proficiency levels as *A2* ($n = 1$), *B1* ($n = 14$), *B2* ($n = 26$), *C1* ($n = 11$), and *C2* ($n = 2$). Figure 3 shows the balanced distribution of participants by English level across conditions. Of all participants, 26 reported prior experience using chatbots in various contexts.

4.2 Experimental setup

The purpose of the experiment is to compare the effects of ICF vs. DCF on language learning and the attitudes of users toward an LLM powered chatbot with open-domain conversation capabilities. A between-subject design was employed, comprising two conditions designed to evaluate the influence of correction timing on language learning and interactional dynamics (ICF vs. DCF as defined in Section 3.1.2). Participants were from four classes (A, B, C, D) and pseudo-randomly assigned to either of the two conditions. This strategy ensured a balanced distribution of participants across the four classes and mitigated potential biases. Among participants who completed the experiment, 24 were in the Immediate condition and 30 in the Delayed condition. This imbalance was due to drop-outs of the study. We analyzed the dropouts and there was no particular pattern among them which could potentially explain the reason for their withdrawal.

4.3 Materials and measures

The chatbot was introduced to participants as "Alex" [3] without specifying gender or age, and positioned as a native English-speaking tutor. Apart from the chatbot system, the materials used in this experiment included two forms distributed to participants before and after the experiment[4]. The objective of these forms was to collect information about the participants and assess changes in their perceptions after engaging in conversations with the chatbot. Additionally, for most questions in the pre-test and post-test, participants were asked to explain the reasoning behind their choices and responses.

4.3.1 Pre-test

Through the pre-test, we aimed to collect participants' information, including socio-demographic data, their First Language (L1), proficiency in English, and prior experience with chatbots and in which contexts. Additionally, we posed four fundamental questions that were repeated in the post-test to assess changes in participant responses after using the chatbot. These questions include:

1. **Timing**: When do you prefer to receive corrections while practicing conversation with a native speaker? (*During the Conversation, After the Conversation, or no correction at all*)
2. **Usage**: would you use a chatbot to practice a language you are learning? (*Yes, No, Maybe*)
3. **Effectiveness**: how effective do you think chatbots are in language education? (7*-point Likert scale*)
4. **Ranking**: among these assistive tools, rank them according to your preference for language practice. (*Robot, Animated Virtual Agent, Language Learning Applications and Websites, Chatbot, Language Book, Gaming, Media*). This was then turned into a score, of 1-7, depending on where they ranked the Chatbot (7 being ranked highest).

4.3.2 Post-test

4.3.2.1 Language assessment

The duration of the experiment may have been insufficient for participants to apply their learned knowledge to the entire English grammar. Consequently, a comprehensive language test was not administered before and after the experiment. Instead, participants' own mistakes served as a preliminary assessment. The participants were given 25 sentences extracted from their utterances and were instructed to identify and correct any mistakes they found. Of the 25 sentences provided, roughly 15 included errors made by the participant, while the remaining sentences were correct. Participants earned one point for each mistake they successfully corrected.

4.3.2.2 Qualitative form

The qualitative form resembled the pre-test and included the same four fundamental questions, along with several supplementary questions about participants' interactions with the chatbot, their perceptions of the chatbot's personality, and their language learning experiences:

- **Gender**: What was Alex's gender? (*(Male, Female, Non Binary, It did not have a gender)*)

[3] Gender neutral name. However, later we learned in Czech it has implication of a male name.

[4] Our repository on Open Science Foundation contains extensive supplementary material, including statistical analysis, graphs, and anonymized data from the experiment (excluding any identifiable or sensitive information about the participants).

- **Age**: How old was Alex?
- **Role**: Who was Alex to you? (*Teacher, Friend, A random person, A mobile application*)
- **Learning**: Do you think you learned anything in English from chatting with Alex? (*Yes, No*)
- **Experience**: How was your experience with the chatbot (*5 point Likert scale*)
- **Politeness**: Was the chatbot polite? (*5 point Likert scale*)

4.3.2.3 Error counts

During the experiment, we recorded the information about all the grammatical mistakes that the users made including the wrong word, whether the correction was provided, and the type of error. This is then used as a measure of language learning outcomes in the analysis.

4.4 Procedure

4.4.1 Consent and Onboarding

The English course instructor mandated the experiment as a required extracurricular activity for students. Only the data of participants who provided informed consent were used in the experiment, and the data of others were deleted. During class, participants completed a pre-test form and were directed to use and sign up for the chatbot system. The experimenter explained the form's questions and remained available to answer any clarifying questions. The initial session involved the chatbot introducing itself to the students and actively participating in open conversation. This allowed participants to identify and report any potential issues to the researchers.

4.4.2 Session schedule and topics

Each chat session was considered complete once the user generated 1,000 characters to ensure consistent chat exposure across participants. Participants completed 12 sessions in total, conducting three per week for 4 weeks (the timeline and sessions shown in the dashboard to the users can be seen in Figure 4). A limit of one session per day was imposed to prevent overwhelming participants with too many sessions in 1 day. Additionally, all sessions were conducted outside of the school environment, allowing participants to work at their own pace. The weekly topics were chosen based on student interest, as determined by pre-experiment voting, and each week focused on a different topic.

4.4.3 Post-session assessment

After finishing their final chat session, participants were directed to the language assessment page. The assessment comprised 25 sentences; each presented individually with a text box allowing participants to provide corrections if they identified any errors. Following the assessment, participants proceeded to the qualitative form. The experiment concluded once both forms were completed.

5 Results and analysis

We used `R 4.3.2` to conduct our statistical testing (R Core Team, 2023). In all of our analysis we considered the conventional $p_{value} < 0.05$ for the significance level.

5.1 User's perception of chatbot's effectiveness for language learning

5.1.1 Composite measure

In **RQ1** and **RQ3** we asked about how effective LLMs can be in language learning. In the pre- and post-questionnaires (see Section 4.3.1), we asked the users to express their opinion about how effective they think chatbots can be (Effectiveness) and how they would rank chatbots compared to other assistive technologies when it comes to language learning (Ranking). We combined the responses by taking the average of these two scores (which were both from 1 to 7, where higher is better).

5.1.2 Pre–post change

The Sign test was conducted to evaluate whether there was a statistically significant change in the Likert scale assessments of effectiveness before and after an intervention, using paired data from 54 participants. Out of the 31 non-zero changes, 22 displayed an increase in effectiveness ratings. The exact binomial test yielded $p_{value} = 0.0294$, suggesting that the probability of observing this level of positive change, or more extreme, by random chance is approximately 2.94% under the null hypothesis that there is no true effect (i.e., a 50% chance of a positive change). This significant result indicates that the effectiveness ratings improved after the intervention. The estimated probability of a positive change (or 'success') is about 70.97%, with a 95% confidence interval ranging from 51.96% to 85.78%. Consequently, these findings support the idea that the intervention had a positive effect on perceived effectiveness of chatbots for language learning.

5.1.3 Between-condition comparisons

Additionally, in **RQ3** we asked about the impact of receiving ICF or DCF from a chatbot on learners' perceptions of chatbot use for language learning. Based on prior research, we posed the question regarding the differential effects of these feedback types on learners' perspectives. We evaluated the normality of the pre- and post-test effectiveness scores for both the ICF and DCF groups utilizing the Shapiro–Wilk test. The results of the Shapiro-Wilk test for the ICF group were $W = 0.978$ ($p = 0.849$) for the post-test and $W = 0.922$ ($p = 0.072$) for the pre-test, suggesting that the scores from both measurements do not significantly deviate from a normal distribution. Similarly, for the DCF group, the Shapiro-Wilk test yielded $W = 0.961$ ($p = 0.334$) for the post-test and $W = 0.964$ ($p = 0.393$) for the pre-test, indicating no significant departure from normality. Therefore, there is insufficient evidence to reject the null hypothesis of normality.

Given these results, the application of parametric tests, such as the independent samples t-test, is justified for comparing the mean

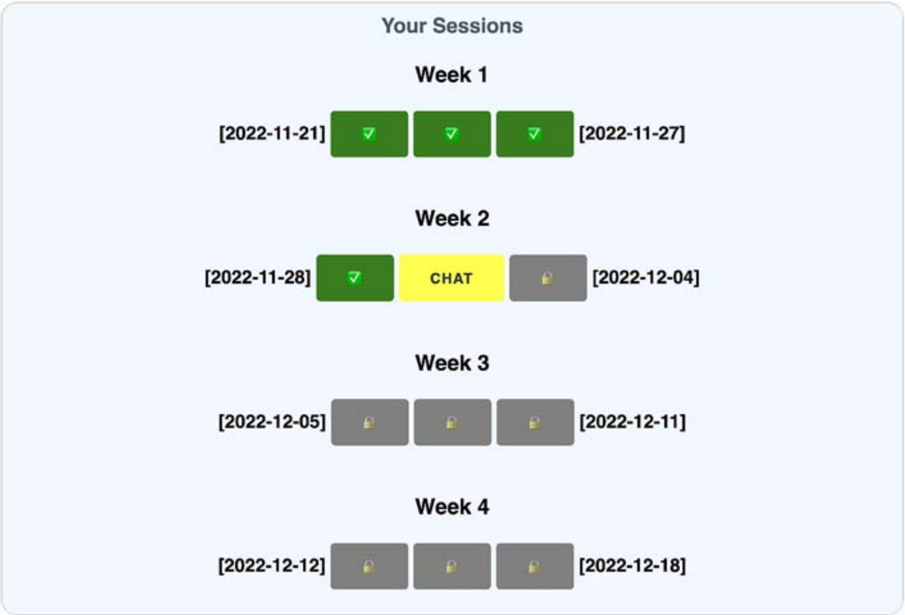


FIGURE 4
The dashboard showing the timeline and sessions to the users.

effectiveness scores between the pre-test and post-test within each group. This decision is supported by the fact that Likert scale data, when aggregated across participants, can produce distributions that approximate normality. Moreover, the *t*-test is robust to deviations from normality, especially with similar sample sizes and when the departure from normality is not severe. Therefore, the preconditions for conducting t-tests on this Likert scale data are satisfied, lending credibility to the subsequent analyses and findings.

Using an independent samples *t*-test, we compared the responses between the two groups from the pre-questionnaire, which showed no significant difference ($t(44.28) = 0.25$, $p = 0.8$, Cohen's $d = 0.07$). However, in the post-questionnaire, there was a significant difference between the mean scores of participants in the ICF group ($\bar{x} = 3.65$) and the DCF group ($\bar{x} = 3.08$), ($t(47.96) = 2.01$, $p = 0.0495$, Cohen's $d = 0.56$). We also conducted two paired *t*-tests for the change in effectiveness scores for the ICF and DCF groups. The ICF group demonstrated a significant increase ($\bar{x} = 1.48$), ($t(22) = 2.63$, $p = 0.015$, Cohen's $d = 0.68$) while the DCF group showed no significant change ($\bar{x} = 0.43$), ($t(29) = 1.15$, $p = 0.259$, Cohen's $d = 0.23$) in effectiveness scores. These findings suggest that the interaction with the ICF chatbot had a slightly higher impact on participants' preference score changes, compared to the DCF chatbot.

5.2 Perception of chatbot's personality

5.2.1 Gender

In the post-test, we asked participants about the chatbot's personality. 56% of the participants considered the chatbot Male. Common reasoning included the name of the chatbot being male, feeling like the conversation had a masculine vibe, and the chatbot answering questions in a way that felt like it would be a male response. A few participants chose a female gender due to the chatbot's name being unisex and feeling like they were getting female responses. Others decided to assign the chatbot no gender (31%) or chose a non-binary gender. Reasons for why this was their choice varied, including not knowing which gender to assign, the chatbot not specifying a gender for itself, or thinking that it was pointless to assign the chatbot a gender.

5.2.2 Age

We asked participants to assign an age to the chatbot. They reported the chatbot's age ($\bar{x} = 30.94$, $s = 12.59$). The majority of participants assigned an age of 30 years. The reasoning behind this is largely associated with the maturity and conversational responses of the chatbot, which was seen as neither too childlike nor too serious. There were some exceptions to this, with participants assigned younger and older ages. Participants who assigned younger ages (generally between 20 and 25) indicated that the chatbot knew a lot about modern things, had conversations and hobbies typical for young people, or cited similarities between themselves and the chatbot. In contrast, those that assigned older ages to the chatbot (between 35 and 70) stated that the chatbot seemed more mature and intelligent, had life experiences, or knew many languages and had higher education.

5.2.3 Social role

We asked the participants who the chatbot was to them in these 4 weeks, in terms of its social role. Although most of the participants used their smartphones to have the chats, only 21.8%

chose “A mobile application”. The rest of the participants chose the other anthropomorphic options: “A random person” (40%), “Friend” (27.3%), and “Teacher” (10.9%).

5.3 Language learning

5.3.1 Subjective measure

To explore **RQ1** and **RQ2**, we asked the participants in the post-test form whether they think they learned anything in English through interactions with the chatbot. It was observed that 72.22% (39 out of 54) reported learning (”Yes”). A one-sample proportions test with continuity correction was conducted to test if this result significantly differed from a 50% chance level. The test yielded a statistically significant deviation ($\chi^2(1) = 9.79$, $p = 0.0017$) from chance level, with the 95% confidence interval for the proportion of positive responses being $58.14\% - 83.14\%$. These results suggest that a substantial majority of the participants perceived the engagement with the chatbot as an educational experience in learning English, supporting the chatbot’s efficacy in language learning.

5.3.2 Objective measure

5.3.2.1 Language assessment

To evaluate language learning outcomes as an answer to **RQ4** and **RQ1**, we compared the final language assessment results and the count of grammatical errors and their types across all sessions between participants who received ICF and DCF. Participants in both groups could successfully spot and correct $\bar{x} = 2.62$ ($n = 53$, $s = 2.6$) of the 15 mistakes shown to them (one participant’s assessment data was excluded due to a technical issue). We explored the possible differences in the mean language assessment scores between the two groups: ICF and DCF. The results of an independent samples t-test showed no significant difference between ICF ($\bar{x} = 3.14$, $s = 2.66$) and DCF ($\bar{x} = 2.18$, $s = 2.51$), ($t(47.99) = 1.35$, $p = 0.092$, Cohen’s $d = 0.37$), indicating that there was no statistically significant difference in language test performance between the two groups.

5.3.2.2 Error counts

We categorized all the error types made by the participants based on the classifications specified in Bryant et al. (2017). The number of grammatical errors vastly differed across different categories with determiners being the most common mistake type. Figure 5 shows the statistics of the error counts in the different categories.

We used the reduction in the number of grammatical errors made by each participant throughout the sessions as a measure of improvement in grammatical knowledge. However, this method may be influenced by factors other than the conditions of the experiment. One such factor is the cumulative number of CFs received by each participant in each session and the specific type of error being addressed. For instance, if a participant received more corrections related to determiners, we would expect them to show a decreasing trend in error counts in subsequent sessions. Another factor is the session itself, as each session covered a different topic. We employed a Generalized Linear Mixed Model (GLMM) to evaluate language learning based on error counts. The model considered the error counts of each session as the response variable, with the condition, session numbers, and error types as fixed effects, and the sessions, participants, and error types as random effects. After fitting the model, the GLMM did not yield a significant relationship between the count of errors and these factors.

5.4 Participant’s experience

Based on the participants’ responses on a scale of 1:Terrible to 5:Amazing, they had a moderately positive experience ($\bar{x} = 3.69$, $s = 0.72$). Also, they considered the chatbot to be polite ($\bar{x} = 4.52$, $s = 0.75$) on the scale 1:Rude to 5:Polite.

We also asked the participants to describe the best and worst parts of their experience. Overall, many of the participants stated that they enjoyed “learning new things” and “having conversations with the chatbot”. They praised the chatbot for understanding jokes, recognizing their mistakes, and having interesting conversations. The participants felt that the chatbot was friendly and attentive. Some of the worst experiences reported by the participants were incomplete sentences, bugs in the user interface, and being asked the same question multiple times. Participants in the immediate condition reported enjoying discussing interesting topics with the chatbot, as well as being impressed with the chatbot’s ability to actively make conversation and ask questions. However, they also experienced negative aspects, such as the chatbot not finishing some sentences and cycling back to asking the same questions. Participants in the delayed condition enjoyed talking about topics they usually do not discuss with other people and also praised the chatbot for accurately and appropriately answering their questions. However, some of their worst experiences include the chatbot writing incomplete sentences, the chatbot wanting to end the session too early, and having to reopen a session when there was a glitch in the system.

5.5 Dropouts in the study

Out of the 66 participants who enrolled in the study, 12 did not complete the experiment. This constitutes a large proportion of participants (of all 66); thus, their data was analyzed separately to identify any underlying systematic factors that led to their non-completion. Three participants completed the pre-test but did not register for the chatbot system, while two completed the pre-test and chatbot registration but did not complete the introductory session. A further two participants only completed the introductory session, and one participant was excluded due to attempting to input copy-pasted text. Three of the remaining four participants were in the Immediate condition and completed 11, 11, and 9 sessions. The final participant was in the Delayed condition and only completed four sessions.

6 Discussion

This study investigated the effects of utilizing LLM-powered chatbots on language learning, emphasizing the role

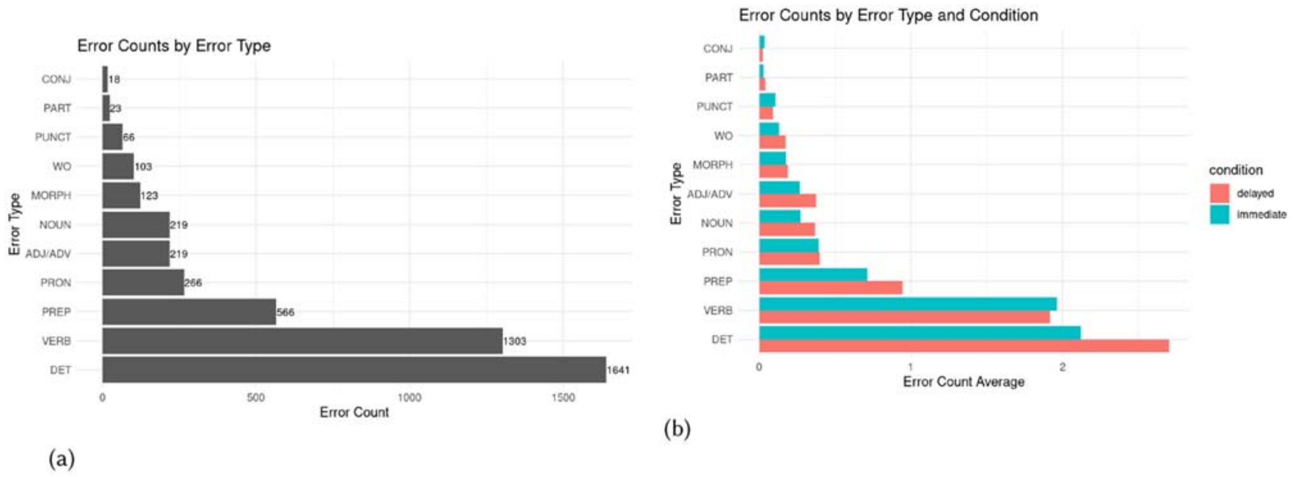


FIGURE 5
(a) Count of all the errors made by all the participants. **(b)** Average number of errors made of each type per participant, per chat session in each condition. Statistics of the error counts.

of CF timing on learners' grammar acquisition and their perceptions of chatbot efficacy. Our research sought to unveil how advancements in LLM technologies could be harnessed to enhance the language learning experience, particularly through real-time CF. By conducting this study, we aimed to validate the potential of LLM-powered conversational systems in educational settings and proposed a framework for future investigations into different CF strategies, including personalization and feedback timing. This endeavor contributes to the broader discussion on optimizing technology-assisted language learning methods.

The prominence of recent generative models like LLMs in the tech landscape inspired our study. Such models open new possibilities for language education, notably the provision of, automatic ICF on learners' grammatical errors–an approach shown to be beneficial for second-language acquisition. Our interest lied in exploring the capability of LLM-powered chatbots to offer this immediate feedback during open-domain conversations, an application not previously possible before the advent of advanced LLMs. This study aimed to bridge the gap in language learning methodologies by leveraging the potential of LLMs for offering real-time, personalized CF, thus paving the way for pioneering the use of chatbots in language education.

The empirical investigation underpinning this study sought to explore several research questions, each contributing valuable insights into the utilization of LLM-powered chatbots for language learning. In addressing **RQ1**, our focus was to understand the role of LLMs in language education, particularly through the lens of providing CF on grammatical errors. The results of the post-questionnaire revealed a positive perception among users regarding the chatbot's effectiveness for language learning in both experimental conditions, underscoring the potential of LLM-powered chatbots in facilitating language acquisition. This was also evident when analyzing the qualitative responses to the open questions regarding the participants' experiences with the chatbot. The majority of participants expressed satisfaction with the chatbot's performance, highlighting its potential as a language learning tool.

In **RQ2** we questioned the learners' perception of the chatbot's personality and its anthropomorphic attributes. Nearly 80% of participants reported feeling they were conversing with a human, not a mere mobile application, highlighting the chatbot's capacity to emulate human-like interactions effectively. Despite participants being aware of engaging with AI, the exercise of attributing age and gender to the chatbot further demonstrated its perceived human-likeness, with only a minority contesting the notion of anthropomorphizing chatbot. This aspect of our findings suggests a significant achievement in creating relatable and engaging learning tools.

In exploring **RQ3**, we investigated the impact of ICF vs. DCF on users' inclination toward employing LLM-powered chatbots among other educational tools and in isolation for language practice. The evidence pointed to a significant contrast in attitudes, with the ICF condition fostering a more favorable perception of chatbots as effective language learning tools compared to the DCF group. This outcome suggests that the timing of feedback plays a crucial role in enhancing learners' engagement and perceived benefits of such technological interventions. The increased attention in turn could indirectly lead to more effective language learning outcomes, as learners are more likely to engage with the chatbot for longer periods and more frequently.

Delineated in **RQ4**, we aimed to assess how CF timing influences grammar learning gains. Contrary to prior studies advocating for the superiority of ICF, our findings did not unveil significant differences in grammar acquisition between the feedback timing conditions, challenging established narratives in the literature. The absence of detectable learning gains might be attributed to the limitations of our experimental design or the methods utilized for measuring learning outcomes. However, it's worth noting the possibility that the insufficient number of participants and the nature of the learning assessments might have influenced these results. Although a tendency toward superior performance in grammar learning was noted within the ICF group compared to the DCF group, the statistical insignificance bars us from drawing definitive conclusions. Future experiments, ideally featuring more controlled designs and larger sample sizes, could shed further light on this matter, as preliminary data hint at an underlying positive trend that warrants deeper investigation.

The granular analysis of grammatical errors revealed insightful patterns, notably the predominance of determiner errors, which offers a focused area for improving personalized feedback mechanisms in language education. The meticulous logging and analysis of such errors not only facilitate a nuanced understanding of common stumbling aspects of certain L2 for learners but also has the potential to transform LLM-powered learning systems into highly adaptive platforms. By tailoring the learning experience to address an individual's specific weaknesses, such technologies stand to significantly enhance the efficacy and personalization of language learning.

6.1 Implications for language learning, conversational systems, and corrective feedback

The efficacy of ICF for language learning, as extensively discussed in our literature review, sets a foundational premise for this study. While the benefits of ICF are well-documented, its provision often demands significant human effort, time, and dedication–resources that are not universally accessible. Personalized feedback, although beneficial for enhancing language proficiency, faces similar constraints within conventional educational settings, where the scope for individualized attention is limited. This predicament underscores the importance of exploring alternative means to deliver ICF efficiently and at scale.

Our investigation reveals that LLM-powered chatbots represent a viable solution to this challenge, enabling the automatic provision of ICF to language learners. Notably, participants receiving ICF via these chatbots exhibited a higher propensity to engage with and recognize the utility of such systems for language learning. This observation underscores the potential of conversational systems to augment traditional language education by offering tailored and immediate feedback, a feature that elevates the learning experience significantly.

The broader implications of our findings extend to the realm of technological interventions in language education. The positive shift in users' attitudes toward language learning, catalyzed by interactions with LLM-powered chatbots, emphasizes the role of technology in fostering engagement and sustained interest in language acquisition. This aligns with contemporary research highlighting the advantages of digital tools in providing interactive, practice-oriented, and personalized learning experiences. Furthermore, the anthropomorphic traits attributed to chatbots by users signal an emerging preference for more human-like interactions within digital learning environments, pointing to design considerations that could enhance the efficacy and appeal of these tools.

The Noticing Hypothesis, integral to our theoretical framework, posits that language acquisition is contingent upon the conscious awareness of linguistic features (Schmidt, 2001, 1990). This hypothesis aligns closely with the operational mechanisms of LLM-powered chatbots employed in our study, especially through the provision of ICF. The chats not only facilitate interaction in the target language but also draw attention to linguistic forms that learners may otherwise overlook. CF, especially when delivered immediately, becomes a pivotal intervention, enabling learners to identify and reconcile discrepancies between their interlanguage and the target language norm. This process is reflective of Schmidt's contention that attention and the resultant noticing are crucial for language learning.

In the context of our findings, the capacity of LLM-powered chatbots to provide ICF resonates with the theoretical underpinnings of the Noticing Hypothesis. By creating a dynamic learning environment where errors are explicitly highlighted, these digital aids effectively foster the learners' ability to notice grammatical pitfalls. This immediate form of feedback not only caters to the learners' need for corrective insight but also aligns with interactionist perspectives that emphasize the role of negative evidence in learning. Contrary to traditional settings where such personalized CF might be scarce, the deployment of chatbots for this purpose represents a significant leap toward facilitating focused language acquisition.

Moreover, our exploration into the nuanced effects of feedback timing on language learning outcomes–an area that reveals complex learner preferences–might also be contemplated through the lens of noticing. While the acquisition of language proficiency involves a multitude of cognitive processes, the engagement with and the subsequent noticing of linguistic forms instigated by ICF could elucidate why learner preferences tend toward immediate feedback. This preference, albeit not directly translating to superior learning gains in our study, underscores the role of chatbots in enhancing engagement and fostering a conducive environment for language acquisition.

The apparent discrepancy between participants' preference for ICF and the non-significant difference in measured learning gains warrants theoretical interpretation. Drawing on our theoretical framework, we propose that ICF and DCF may differentially influence affective vs. cognitive outcomes. The Noticing Hypothesis (Schmidt, 1990) suggests that ICF draws immediate attention to linguistic forms, which our preference data support—learners valued this real-time salience. However, in our open-domain conversational design, neither condition systematically pushed learners to reproduce corrected forms: ICF was followed by continued conversation on new topics, while DCF was presented after the interaction concluded. This absence of structured output practice—a key mechanism in the Comprehensible Output Hypothesis Swain (2005)—may explain why both conditions yielded comparable cognitive gains despite differing in perceived salience.

Furthermore, the cognitive comparison and reactivation theories discussed in our literature review (Section 2.2) offer additional perspective. While ICF enables immediate comparison between erroneous and target forms, DCF may trigger deeper reactivation of the original context, engaging different memory consolidation processes. In ecologically valid, open-domain conversations—as opposed to controlled experiments targeting isolated grammatical structures—these mechanisms may operate in parallel, yielding comparable learning outcomes despite divergent learner preferences.

The human–AI interaction dynamics noted in our Introduction may also contribute to these findings. Even when chatbot feedback quality matches human-delivered correction, learners' awareness of interacting with an artificial agent could alter how they process and internalize CF. The strong anthropomorphization observed in our study (approximately 80% of participants attributed human-like characteristics to the chatbot) suggests that learners engaged socially with the system, potentially mediating feedback effectiveness in ways that differ from human–human interaction contexts studied in prior CF research.

Finally, statistical power considerations merit acknowledgment. Although our sample size ($N = 54$) was sufficient to detect differences in preference, it may have been underpowered to detect subtle differences in grammar acquisition, particularly given the variability inherent in open-domain conversation data. The observed trend favoring the ICF group in language assessment scores, while not reaching statistical significance, suggests that larger-scale studies with more controlled designs may reveal effects that our ecological approach could not detect definitively.

While our results challenge the presumed necessity of ICF for effective language learning, it is crucial to approach these findings with caution. The unique methodological setup of our study differs significantly from controlled experiments focusing on specific grammatical structures, suggesting that our insights into feedback timing necessitate further validation under varied conditions.

The integration of LLM technologies in language learning introduces a promising avenue for evolving pedagogical strategies, especially within conversational systems. The advent of LLMs enables a more dynamic application of immediate and personalized feedback, previously constrained by technological limitations. This evolution not only capitalizes on the documented benefits of ICF but also opens exploratory pathways to reassess the potential of DCF within the broader landscape of technology-assisted language learning. Consequently, these developments contribute to a more nuanced understanding of how digital tools can be optimized for language education, aligning technological capabilities with pedagogical objectives to enhance learner outcomes and experiences effectively.

While AI-driven language learning technologies are advancing rapidly, they are not yet at a stage where they can replace human educators. However, as technology continues to evolve, it is possible that future advancements may significantly transform how knowledge is shared and transferred across generations. Given the current state of AI, it is crucial to approach these tools as assistive, rather than as alternatives to well-established language learning methods. Studies like ours should not be interpreted as advocating for the replacement of proven educational approaches but rather as contributing to the development of modern, supplementary tools that can enhance the language learning experience. Thus, as research progresses, ensuring that these tools are designed with a clear pedagogical purpose will be essential. Their role should be considered within a broader educational framework that prioritizes human-guided instruction while leveraging AI to provide scalable and adaptive learning opportunities.

6.2 Limitations and future work

Our study marks a critical step toward understanding the roles of conversational systems and CF in educational contexts. Nonetheless, it showcases limitations that inevitably pave the way for future research directions. Primarily, our approach to measuring language learning gains, specifically the reliance on participants' grammatical mistakes without a pre-test for comparison, presents challenges in conclusively capturing learning gains. Additionally, self-reported proficiency levels may introduce bias, as learners might over- or underestimate their abilities, though our pseudo-random assignment helped mitigate imbalances. The varying chat session topics and the influence of external factors such as participants' tiredness or mood potentially affected the consistency of error counts or the users' attention to the CFs, underscoring the need for more controlled experimental designs in future research.

Additionally, time was not controlled as a variable in our study due to several influencing factors, such as the platform participants used (e.g., mobile or computer), their typing speed, time of day, and the varying cognitive demands of different conversation topics. For example, participants may have needed more time to answer complex questions requiring reflection, such as "Where do you want to travel to?" compared to simpler questions like "Where have you traveled to?" These differences in response time, much like in human interactions, make it difficult to draw reliable comparisons based on time alone. This was a conscious design choice, where the time taken to reach 1000 characters was not considered a variable, and thus, we argue this is not necessarily a limitation but rather an alternative design approach.

Furthermore, the absence of a control group, dictated by our study's focus on comparing ICF and DCF conditions, constrains the breadth of our findings. The Sign test's reliance on non-zero changes might not fully capture the nuances of user perception shifts. Also, the large dropout rate calls for a closer examination of participant engagement strategies and the user interface's role in maintaining interest and motivation over time. Moreover, the generalization of study findings to larger populations would benefit from a larger sample size to mitigate variability and strengthen the study's external validity. Additionally, taking into account user characteristics when designing the chatbot's persona could make the tool more relatable and compelling for different learners.

Another limitation of this study is the methodology of asking participants to correct errors they previously made in their own writing. While there may have been some priming effect, it is likely minimal, as the sentences were spread out over a month, making it difficult for participants to remember them. Additionally, although the experiment lasted for a month, this duration and the type of activity may not be sufficient to generalize the learning outcomes across all aspects of language acquisition. Ideally, a comprehensive pre-test and post-test would have provided a more reliable measure of learning effects. However, due to logistical constraints and the nature of the study, this was not feasible. Asking participants to correct their own mistakes acted as a mitigation for the lack of a pre-test, as it can be inferred that if participants made an error once, they did not know the correct form at that time, which indirectly serves as a pre-test.

A more rigorous and controlled experimental setup, perhaps focusing on specific grammatical structures like determiners, could enhance the reliability of findings related to CF timing effects. Future studies might benefit from incorporating a control group and a within-subject design to furnish a more nuanced understanding of feedback timing on individual learning trajectories. Moreover, addressing the high dropout rate experienced in our study necessitates a closer examination of participant engagement strategies and the chatbot's user interface design.

The nuanced effects observed between ICF and DCF, which seemed to influence learner preferences rather than direct language proficiency, call for further investigation into the integration of CF mechanisms within conversational systems. Technological advancements since our study's inception, particularly the evolution of LLM technologies beyond early versions of GPT-3, present opportunities to reassess the study's findings with more sophisticated models that could yield different results. Additionally, future work could explore a unified design approach where LLM seamlessly integrates GEC with conversational functionalities, potentially enhancing system coherence and reducing error rates.

Extending research to include participants from diverse backgrounds would enable a comparative analysis of the effectiveness of LLM-powered chatbots across different learner profiles. A more thorough comparison of various CF types, as well as a controlled examination of feedback timing (immediate vs. delayed), would deepen our understanding of optimal feedback strategies. Long-term engagement with chatbot systems could illuminate sustained language learning outcomes and user perceptions, inviting further exploration into dialogue strategies, conversation depth, and chatbot personification's impact on learning and engagement.

In sum, while this study contributes valuable insights into the application of LLM-powered chatbots for language learning, the landscape of conversational systems and CF is ripe for further exploration. The advent of advanced LLM technologies and a growing emphasis on personalized, technology-assisted learning underscore the need for continued research in this domain. By addressing the highlighted limitations and pursuing the outlined future directions, subsequent studies can build on our findings to optimize the educational value of chatbot systems in language learning contexts.

6.3 Conclusion

This study has provided evidence LLM-powered chatbots can be utilized effectively for language learning, with positive changes in users' perceptions of their effectiveness. While the timing of Corrective Feedback (CF) (immediate vs. delayed) did not show a significant impact on language learning outcomes, it did influence user preference, indicating a nuanced role of feedback in the learning process. Participants' personalization of the chatbot suggests potential for enhancing engagement through human-like interactions. Although technical issues and a high dropout rate indicate areas for improvement, the general participant experience was positive. These results add to the understanding of Large Language Models' potential in language education and highlight the importance of carefully designing chatbot-led interventions to align with learner characteristics and preferences. Going forward, it is clear that while chatbots hold potential as educational tools, optimizing their use for individualized language learning remains an area ripe for further research.

Data availability statement

The datasets presented in this study can be found in online repositories. The data and analysis can be found at: OSF (https://osf.io/m9qgf/) and the code of the web-app of the project can be accessed at: GitHub (https://ali.mk/ChatBot2023).

Ethics statement

Ethical approval was not required for the studies involving humans because according to the local regulations of the research ethics, the type of research conducted by this study, does not require an ethical approval, as there is no effect on participants, no any identifiable data is shared, and no processing is done on any sensitive data. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

AM: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. BT: Data curation, Formal analysis, Methodology, Software, Writing – original draft, Writing – review & editing. ML: Software, Writing – original draft. GS: Funding acquisition, Supervision, Writing – review & editing.

Funding

The author(s) declared that financial support was received for this work and/or its publication. This project has received funding from the European Union's Horizon 2020 re-search and innovation programme under the Marie Sklodowska-Curie Actions grant agreement No. 857897 (https://doi.org/10.3030/857897.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author GS declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Alkhazraji, A. M. (2018). Analyzing the impact of teacher talk on english grammar learning: with correlation to the procedures in classroom interaction. *J. Lang. Teach. Res*. 9:1109. doi: 10.17507/jltr.0905.27

Ammar, A., and Hassan, R. M. (2018). Talking it through: collaborative dialogue and second language learning. *Lang. Learn*. 68, 46–82. doi: 10.1111/lang.12254

Atkinson, R. C., and Shiffrin, R. M. (1968). "Human memory: a proposed system and its control processes," in *Psychology of Learning and Motivation*, vol. 2, eds. K. W. Spence and J. T. Spence (Academic Press), 89–195. doi: 10.1016/S0079-7421(08)60422-3

Ayedoun, E., Hayashi, Y., and Seta, K. (2015). A conversational agent to encourage willingness to communicate in the context of English as a foreign language. *Proc. Comput. Sci*. 60, 1433–1442. doi: 10.1016/j.procs.2015.08.219

Ayedoun, E., Hayashi, Y., and Seta, K. (2019). Adding communicative and affective strategies to an embodied conversational agent to enhance second language learners' willingness to communicate. *Int. J. Artif. Intell. Educ*. 29, 29–57. doi: 10.1007/s40593-018-0171-6

Baddeley, A. (2000). The episodic buffer: a new component of working memory? *Trends Cogn. Sci*. 4, 417–423. doi: 10.1016/S1364-6613(00)01538-2

Baddeley, A. D., Hitch, G. J., and Allen, R. J. (2009). Working memory and binding in sentence recall. *J. Mem. Lang*. 61, 438–456. doi: 10.1016/j.jml.2009.05.004

Bao, M. (2019). Can home use of speech-enabled artificial intelligence mitigate foreign language anxiety - investigation of a concept. *Arab. World Engl. J*. 28–40. doi: 10.24093/awej/call5.3

Baz, E. H., Balçıkanlı, C., and Cephe, P. T. (2016). Perceptions of English instructors and learners about corrective feedback. *Eur. J. For. Lang. Teach*. 1. doi: 10.46827/ejfl.v0i0.331

Beuningen, C. G., v., Jong, N. H., d., and Kuiken, F. (2008). The effect of direct and indirect corrective feedback on L2 learners' written accuracy. *Int. J. Appl. Linguist*. 156, 279–296. doi: 10.2143/ITL.156.0.2034439

Bibauw, S., François, T., and Desmet, P. (2019). Discussing with a computer to practice a foreign language: research synthesis and conceptual framework of dialogue-based CALL. *Comput. Assis. Lang. Learn*. 32, 827–877. doi: 10.1080/09588221.2018.1535508

Bitchener, J. (2008). Evidence in support of written corrective feedback. *J. Sec. Lang. Writ*. 17, 102–118. doi: 10.1016/j.jslw.2007.11.004

Brown, D., Liu, Q., and Norouzian, R. (2023). Effectiveness of written corrective feedback in developing L2 accuracy: a Bayesian meta-analysis. *Lang. Teach. Res*. 13621688221147374. doi: 10.1177/13621688221147374

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). Language models are few-shot learners. *arXiv:2005.14165 [cs] version: 4*. doi: 10.48550/arXiv.2005.14165

Bryant, C., Felice, M., and Briscoe, T. (2017). "Automatic annotation and evaluation of error types for grammatical error correction," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Vancouver, BC: Association for Computational Linguistics), 793–805. doi: 10.18653/v1/P17-1074

Campbell, R. N. (1970). An evaluation and comparison of present methods for teaching english gammar to speakers of other languages. *TESOL Quart*. 4:37. doi: 10.2307/3585777

Canals, L., Granena, G., Yilmaz, Y., and Malicka, A. (2021). The relative effectiveness of immediate and delayed corrective feedback in video-based computer-mediated communication. *Lang. Teach. Res*. 29, 242–268. doi: 10.1177/13621688211052793

Candel, C., Vidal-Abarca, E., Cerdán, R., Lippmann, M., and Narciss, S. (2020). Effects of timing of formative feedback in computer-assisted learning environments. *J. Comput. Assis. Learn*. 36, 718–728. doi: 10.1111/jcal.12439

Chase, C. C., Chin, D. B., Oppezzo, M. A., and Schwartz, D. L. (2009). Teachable agents and the protégé effect: Increasing the effort towards learning. *J. Sci. Educ. Technol*. 18, 334–352. doi: 10.1007/s10956-009-9180-4

Chen, W. (2017). The effect of conversation engagement on L2 learning opportunities. *ELT J*. 71, 329–340. doi: 10.1093/elt/ccw075

Cheng, X., and Xu, J. (2025). Engaging second language (L2) students with synchronous written corrective feedback in technology-enhanced learning contexts: a mixed-methods study. *Hum. Soc. Sci. Commun*. 12:712. doi: 10.1057/s41599-025-05007-3

Coniam, D. (2014). The linguistic accuracy of chatbots: usability from an ESL perspective. *Text Talk* 34:18. doi: 10.1515/text-2014-0018

Cornillie, F., Clarebout, G., and Desmet, P. (2012). Between learning and playing? Exploring learners' perceptions of corrective feedback in an immersive game for English pragmatics. *ReCALL* 24, 257–278. doi: 10.1017/S0958344012000146

De Gasperis, G., and Florio, N. (2012). "Learning to read/type a second language in a chatbot enhanced environment," in *International Workshop on Evidence-Based Technology Enhanced Learning*, vol. 152, eds. P. Vittorini, R. Gennari, I. Marenzi, F. de la Prieta, J. M. C., and Rodríguez (Berlin; Heidelberg: Springer Berlin Heidelberg), 47–56. doi: 10.1007/978-3-642-28801-2_6

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805 [cs]*. doi: 10.48550/arXiv.1810.04805

Dina, A., and Ciornei, S.-I. (2013). The advantages and disadvantages of computer assisted language learning and teaching for foreign languages. *Proc. Soc. Behav. Sci*. 76, 248–252. doi: 10.1016/j.sbspro.2013.04.107

Dokukina, I., and Gumanova, J. (2020). The rise of chatbots - new personal assistants in foreign language learning. *Proc. Comput. Sci*. 169, 542–546. doi: 10.1016/j.procs.2020.02.212

Dörnyei, Z. (2007). Research Methods in Applied Linguistics: Quantitative, Qualitative, and Mixed Methodologies. Oxford Applied Linguistics. Oxford; New York, NY: Oxford University Press.

Ellis, R. (2009). A typology of written corrective feedback types. *ELT J*. 63, 97–107. doi: 10.1093/elt/ccn023

Ellis, R. (2011). "Corrective feedback in language teaching," in *Handbook of Research in Second Language Teaching and Learning*, ed. H. Eli (New York, NY: Routledge), 593–610.

Ellis, R. (2021). "Explicit and implicit oral corrective feedback," in *The Cambridge Handbook of Corrective Feedback in Second Language Learning and Teaching, 1st Edn*., eds. H. Nassaji and E. Kartchava (Cambridge: Cambridge University Press), 341–364. doi: 10.1017/9781108589789.017

Ellis, R., Loewen, S., and Erlam, R. (2006). Implicit and explicit corrective feedback and the acquisition of L2 grammar. *Stud. Sec. Lang. Acquis*. 28, 339–368. doi: 10.1017/S0272263106060141

Fryer, L., and Carpenter, R. (2006). Bots as language learning tools. *Lang. Learn. Technol.* 10, 8–14. doi: 10.64152/10125/44068

Fryer, L., Coniam, D., Carpenter, R., and Lăpuşneanu, D. (2020). Bots for language learning now: current and future directions. *Lang. Learn. Technol.* 24, 8–22. doi: 10.64152/10125/44719

Fryer, L. K., Nakao, K., and Thompson, A. (2019). Chatbot learning partners: connecting learning experiences, interest and competence. *Comput. Hum. Behav.* 93, 279–289. doi: 10.1016/j.chb.2018.12.023

Fu, M., and Li, S. (2019). The associations between individual differences in working memory and the effectiveness of immediate and delayed corrective feedback. *J. Sec. Lang. Stud.* 2, 233–257. doi: 10.1075/jsls.19002.fu

Fu, M., and Li, S. (2022). The effects of immediate and delayed corrective feedback on L2 development. *Stud. Sec. Lang. Acquis.* 44, 2–34. doi: 10.1017/S0272263120000388

Gernsbacher, M. A. (1990). *Language Comprehension As Structure Building*. New York, NY: Psychology Press. doi: 10.21236/ADA221854

Grgurovic, M., Chapelle, C. A., and Shelley, M. C. (2013). A meta-analysis of effectiveness studies on computer technology-supported language learning. *ReCALL* 25, 165–198. doi: 10.1017/S0958344013000013

Haines, S. (1995). For and against: pairwork. *Modern Engl. Teach.* 4, 55–58.

Haristiani, N. (2019). Artificial intelligence (ai) chatbot as language learning medium: an inquiry. *J. Phys.: Conf. Ser.* 1387:012020. doi: 10.1088/1742-6596/1387/1/012020

Haristiani, N., and Rifa'i, M. (2020). Combining chatbot and social media: enhancing personal learning environment (ple) in language learning. *Indonesian J. Sci. Technol.* 5, 487–506. doi: 10.17509/ijost.v5i3.28687

Havranek, G. (2002). When is corrective feedback most likely to succeed? *Int. J. Educ. Res.* 37, 255–270. doi: 10.1016/S0883-0355(03)00004-1

Heift, T. (2004). Corrective feedback and learner uptake in CALL. *ReCALL* 16, 416–431. doi: 10.1017/S0958344004001120

Henderson, M., Bearman, M., Chung, J., Fawns, T., Buckingham Shum, S., Matthews, K. E., et al. (2025). Comparing generative AI and teacher feedback: Student perceptions of usefulness and trustworthiness. *Assess. Eval. High. Educ.* 1–16. doi: 10.1080/02602938.2025.2502582

Horwitz, E. K., Horwitz, M. B., and Cope, J. (1986). Foreign language classroom anxiety. *Modern Lang. J.* 70, 125–132. doi: 10.1111/j.1540-4781.1986.tb05256.x

Huang, W., Hew, K. F., and Fryer, L. K. (2022). Chatbots for language learning-are they really useful? A systematic review of chatbot-supported language learning. *J. Comput. Assisted Learn.* 38, 237–257. doi: 10.1111/jcal.12610

Hymes, D. (1972). "On communicative competence," *Sociolinguistics: Selected Readings*, eds. J. B. Pride, and J. Holmes (Harmondsworth: Penguin), 269–293.

Jacobsen, L. J., Mertens, U., Jansen, T., and Weber, K. E. (2025). AI, expert or peer? – examining the impact of perceived feedback source on pre-service teachers feedback Perception and uptake. *arXiv*. Available online at: https://arxiv.org/abs/2507.16013 (Accessed December 22, 2025).

Jelinski, J. B., and VanPatten, B. (1997). Input processing and grammar instruction in second language acquisition. *Hispania* 80:811. doi: 10.2307/345093

Jia, J., Chen, Y., Ding, Z., and Ruan, M. (2012). Effects of a vocabulary acquisition and assessment system on students' performance in a blended learning class for English subject. *Comput. Educ.* 58, 63–76. doi: 10.1016/j.compedu.2011.08.002

Kaliisa, R., Misiejuk, K., López-Pernas, S., and Saqr, M. (2026). How does artificial intelligence compare to human feedback? A meta-analysis of performance, feedback perception, and learning dispositions. *Educ. Psychol.* 46, 80–111. doi: 10.1080/01443410.2025.2553639

Kang, M. (2018). A study of chatbot personality based on the purposes of chatbot. *J. Korea Contents Assoc.* 18, 319–329. doi: 10.5392/JKCA.2018.18.05.319

Kerlyl, A., Hall, P., and Bull, S. (2007). "Bringing chatbots into education: towards natural language negotiation of open learner models," in *Applications and Innovations in Intelligent Systems XIV*, eds. R. Ellis, T. Allen, and A. Tuson (London: Springer), 179–192. doi: 10.1007/978-1-84628-666-7_14

Kim, N.-Y. (2018a). Chatbots and Korean EFL students' English vocabulary learning. *J. Dig. Converg.* 16, 1–7. doi: 10.14400/JDC.2018.16.2.001

Kim, N.-Y. (2018b). A study on chatbots for developing Korean college students' English listening and reading skills. *J. Dig. Converg.* 16, 19–26. doi: 10.14400/JDC.2018.16.8.019

Kim, N.-Y., Cha, Y., and Kim, H.-S. (2019). Future English learning: chatbots and artificial intelligence. *Multimedia-Assis. Lang. Learn.* 22, 32–53. doi: 10.15702/mall.2019.22.3.32

Kim, Y., Choi, B., Kang, S., Kim, B., and Yun, H. (2020). Comparing the effects of direct and indirect synchronous written corrective feedback: learning outcomes and students' perceptions. *For. Lang. Ann.* 53, 176–199. doi: 10.1111/flan.12443

Krashen, S. (1981). Second language acquisition. *Sec. Lang. Learn.* 3, 19–39. doi: 10.1017/S0272263100004198

Kuhail, M. A., Thomas, J., Alramlawi, S., Shah, S. J. H., and Thornquist, E. (2022). Interacting with a chatbot-based advising system: understanding the effect of chatbot personality and user gender on behavior. *Informatics* 9:81. doi: 10.3390/informatics9040081

Li, H. (2023). A review on corrective feedback research of the recent 20 years. *Int. J. Educ. Hum.* 9, 190–195. doi: 10.54097/ijeh.v9i3.10621

Li, S. (2010). The effectiveness of corrective feedback in SLA: a meta-analysis. *Lang. Learn.* 60, 309–365. doi: 10.1111/j.1467-9922.2010.00561.x

Li, S., Zhu, Y., and Ellis, R. (2016). The effects of the timing of corrective feedback on the acquisition of a new linguistic structure. *Modern Lang. J.* 100, 276–295. doi: 10.1111/modl.12315

Li, Y., Chen, C.-Y., Yu, D., Davidson, S., Hou, R., Yuan, X., et al. (2022). "Using chatbots to teach languages," in *Proceedings of the Ninth ACM Conference on Learning @ Scale, L@S '22* (New York, NY: Association for Computing Machinery), 451–455. doi: 10.1145/3491140.3528329

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., et al. (2019). RoBERTa: a robustly optimized BERT pretraining approach. *arXiv:1907.11692 [cs]*. doi: 10.48550/arXiv.1907.11692

Long, M. H. (1981). Input, interaction, and second-language acquisition. *Ann. N. Y. Acad. Sci.* 379, 259–278. doi: 10.1111/j.1749-6632.1981.tb42014.x

Long, M. H. (1983). Linguistic and conversational adjustments to non-native speakers. *Stud. Sec. Lang. Acquis.* 5, 177–193. doi: 10.1017/S0272263100004848

Long, M. H. (1996). "The role of the linguistic environment in second language acquisition," in *Handbook of Second Language Acquisition* (New York, NY: Academic Press). doi: 10.1016/B978-012589042-7/50015-3

Lyster, R., and Ranta, L. (1997). Corrective feedback and learner uptake: negotiation of form in communicative classrooms. *Stud. Sec. Lang. Acquis.* 19, 37–66. doi: 10.1017/S0272263197001034

Macintyre, P. D. (2007). Willingness to communicate in the second language: understanding the decision to speak as a volitional process. *Modern Lang. J.* 91, 564–576. doi: 10.1111/j.1540-4781.2007.00623.x

Martinussen, R., Hayden, J., Hogg-Johnson, S., and Tannock, R. (2005). A meta-analysis of working memory impairments in children with attention-deficit/hyperactivity disorder. *J. Am. Acad. Child Adolesc. Psychiatry* 44, 377–384. doi: 10.1097/01.chi.0000153228.72591.73

Mehra, B. (2021). Chatbot personality preferences in Global South urban English speakers. *Soc. Sci. Hum. Open* 3:100131. doi: 10.1016/j.ssaho.2021.100131

Metcalfe, J., Kornell, N., and Finn, B. (2009). Delayed versus immediate feedback in children's and adults' vocabulary learning. *Mem. Cogn.* 37, 1077–1087. doi: 10.3758/MC.37.8.1077

Michel, P., Levy, O., and Neubig, G. (2019). Are sixteen heads really better than one? *arXiv:1905.10650 [cs]*. doi: 10.48550/arXiv.1905.10650

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*, vol. 26, eds. C. J. Burges, L. Bottou, M. Welling, Z. Ghahramani, Z., and K. Q. Weinberger (Curran Associates, Inc.). Available online at: https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf (Accessed March 23, 2024).

Nakaya, K., and Murota, M. (2013). Development and evaluation of an interactive english conversation learning system with a mobile device using topics based on the life of the learner. *Res. Pract. Technol. Enhanced Learn.* 8, 65–89. doi: 10.58459/rptel.2013.865-89

Nass, C., and Moon, Y. (2000). Machines and mindlessness: social responses to computers. *J. Soc. Issues* 56, 81–103. doi: 10.1111/0022-4537.00153

Nissen, M., Rüegger, D., Stieger, M., Flückiger, C., Allemand, M., v Wangenheim, F., et al. (2022). The effects of health care chatbot personas with different social roles on the client-chatbot bond and usage intentions: development of a design codebook and web-based study. *J. Med. Internet Res.* 24:e32630. doi: 10.2196/32630

Norris, J. M., and Ortega, L. (2000). Effectiveness of L2 instruction: a research synthesis and quantitative meta-analysis. *Lang. Learn.* 50, 417–528. doi: 10.1111/0023-8333.00136

OpenA, I. (2022). *ChatGPT: Optimizing Language Models for Dialogue*. Available online at: https://openai.com/index/chatgpt/ (Accessed February 1, 2023).

Opitz, B., Ferdinand, N. K., and Mecklinger, A. (2011). Timing matters: the impact of immediate and delayed feedback on artificial language learning. *Front. Hum. Neurosci.* 5:8. doi: 10.3389/fnhum.2011.00008

Penning de Vries, B., Cucchiarini, C., Strik, H., and van Hout, R. (2011). "Adaptive corrective feedback in second language learning," in *Interdisciplinary Approaches to Adaptive Learning. A Look at the Neighbours, Communications in Computer and Information Science*, eds. S. De Wannemacker, G. Clarebout, and P. De Causmaecker (Berlin; Heidelberg: Springer), 1–14. doi: 10.1007/978-3-642-20074-8_1

Pérez, J. Q., Daradoumis, T., and Puig, J. M. M. (2020). Rediscovering the use of chatbots in education: a systematic literature review. *Comput. Applic. Eng. Educ.* 28, 1549–1565. doi: 10.1002/cae.22326

Pham, X. L., Pham, T., Nguyen, Q. M., Nguyen, T. H., and Cao, T. T. H. (2018). "Chatbot as an intelligent personal assistant for mobile language learning," in *Proceedings of the 2018 2nd International Conference on Education and E-Learning* (Bali, Indonesia: ACM), 16–21. doi: 10.1145/3291078.3291115

Philp, J., Walter, S., and Basturkmen, H. (2010). Peer interaction in the foreign language classroom: what factors foster a focus on form? *Lang. Awar.* 19, 261–279. doi: 10.1080/09658416.2010.516831

Pica, T. (1988). Interlanguage adjustments as an outcome of NS-NNS negotiated interaction. *Lang. Learn.* 38, 45–73. doi: 10.1111/j.1467-1770.1988.tb00401.x

Pica, T. (1996). Second language learning through interaction: multiple perspectives. *Educ. Linguist.* 12, 1–22.

Prithivida, D. (2023). *Gramformer*. Available online at: https://github.com/PrithivirajDamodaran/Gramformer (Accessed January 19, 2024).

Quinn, P. G. (2021). "Corrective feedback timing and second language grammatical development: research, theory, and practice," in *The Cambridge Handbook of Corrective Feedback in Second Language Learning and Teaching*, eds. E. Kartchava and H. Nassaji (Cambridge: Cambridge University Press), 322–340. doi: 10.1017/9781108589789.016

Qun, Y. (2025). Investigating the impact of immediate vs. delayed feedback timing on motivation and language learning outcomes in online education: perspectives from feedback intervention theory. *Learn. Motiv.* 90:102132. doi: 10.1016/j.lmot.2025.102132

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog* 1:9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* 21, 1–67. Available online at: http://jmlr.org/papers/v21/20-074.html (Accessed January 19, 2024).

Rassaei, E. (2013). Corrective feedback, learners' perceptions, and second language development. *System* 41, 472–483. doi: 10.1016/j.system.2013.05.002

Rassaei, E. (2023). The interplay between corrective feedback timing and foreign language anxiety in L2 development. *Lang. Teach. Res.* 13621688231195141. doi: 10.1177/13621688231195141

Ruan, S., Jiang, L., Xu, Q., Liu, Z., Davis, G. M., Brunskill, E., et al. (2021). "EnglishBot: an AI-powered conversational system for second language learning," in *26th International Conference on Intelligent User Interfaces, IUI '21* (New York, NY: Association for Computing Machinery), 434–444. doi: 10.1145/3397481.3450648

Ruan, S., Willis, A., Xu, Q., Davis, G. M., Jiang, L., Brunskill, E., et al. (2019). "BookBuddy: turning digital materials into interactive foreign language lessons through a voice chatbot," in *Proceedings of the Sixth (2019) ACM Conference on Learning @ Scale (L@S '19)(Article 30)* (New York, NY: ACM), 1–4. doi: 10.1145/3330430.3333643

Ruane, E., Farrell, S., and Ventresque, A. (2021). "User perception of text-based chatbot personality," in *Chatbot Research and Design, Lecture Notes in Computer Science*, eds. A. Følstad, T. Araujo, S. Papadopoulos, E. L.-C. Law, E. Luger, M. Goodwin, and P. B. Brandtzaeg (Cham: Springer International Publishing), 32–47. doi: 10.1007/978-3-030-68288-0_3

Saito, K., and Akiyama, Y. (2017). Video-based interaction, negotiation for comprehensibility, and second language speech learning: a longitudinal study. *Lang. Learn.* 67, 43–74. doi: 10.1111/lang.12184

Schmidt, R. (2001). "Attention," in *Cognition and Second Language Instruction*, ed. P. Robinson (Cambridge: Cambridge University Press), 3–32. doi: 10.1017/CBO9781139524780.003

Schmidt, R. W. (1990). The role of consciousness in second language learning. *Appl. Ling.* 11, 129–158. doi: 10.1093/applin/11.2.129

Sheen, Y. (2010). Differential effects of oral and written corrective feedback in the esl classroom. *Stud. Sec. Lang. Acquis.* 32, 203–234. doi: 10.1017/S0272263109990507

Shintani, N., and Aubrey, S. (2016). The effectiveness of synchronous and asynchronous written corrective feedback on grammatical accuracy in a computer-mediated environment. *Modern Lang. J.* 100, 296–319. doi: 10.1111/modl.12317

Shumanov, M., and Johnson, L. (2021). Making conversations with chatbots more personalized. *Comput. Hum. Behav.* 117:106627. doi: 10.1016/j.chb.2020.106627

Soyoof, A., Reynolds, B. L., Rassaei, E., Kao, C.-W., and Van Ha, X. (2025). From teachers to chatbots: scaffolded corrective feedback and student trust in online L2 English classrooms. *Comput. Educ.: Artif. Intell.* 10:100530. doi: 10.1016/j.caeai.2025.100530

Storch, N., and Wigglesworth, G. (2010). Learners' processing, uptake, and retention of corrective feedback on writing: case studies. *Stud. Sec. Lang. Acquis.* 32, 303–334. doi: 10.1017/S0272263109990532

Swain, M. (1985). Communicative competence: some roles of comprehensible input and comprehensible output in its development. *Input Sec. Lang. Acquis.* 15, 165–179.

Swain, M. (2005). "The output hypothesis: theory and research," in *Handbook of Research in Second Language Teaching and Learning*, ed. H. Eli (Mahwah, NJ: Lawrence Erlbaum Associates), 471–483. doi: 10.4324/9781410612700-34

Swain, M., and Suzuki, W. (2008). "Interaction, output, and communicative language learning," in *The Handbook of Educational Linguistics*, eds. B. Spolsky, and F. M. Hult (Carlton, VIC: John Wiley & Sons, Ltd.), 557–570. doi: 10.1002/9780470694138.ch39

Tan, X., Reynolds, B. L., and Ha, X. V. (2022). Oral corrective feedback on lexical errors: a systematic review. *Appl. Ling. Rev.* 15, 1177–1221. doi: 10.1515/applirev-2022-0053

Truscott, J. (2007). The effect of error correction on learners' ability to write accurately. *J. Sec. Lang. Writ.* 16, 255–272. doi: 10.1016/j.jslw.2007.06.003

Tu, J. (2020). Learn to speak like a native: ai-powered chatbot simulating natural conversation for language tutoring. *J. Phys.: Conf. Ser.* 1693:012216. doi: 10.1088/1742-6596/1693/1/012216

Vygotsky, L. S. (1978). *Mind in Society: Development of Higher Psychological Processes*. Harvard University Press. Available online at: http://www.jstor.org/stable/j.ctvjf9vz4 (Accessed February 5, 2026).

Wahyuni, N., and Afrianti, I. (2021). The contribution of speaking practice with the native speaker to student's speaking ability in junior high school. *Ainara J.* 2, 247–252. doi: 10.54371/ainj.v2i3.88

Wang, D. (2024). Teacher- versus AI-generated (poe application) corrective feedback and language learners' writing anxiety, complexity, fluency, and accuracy. *Int. Rev. Res. Open Distrib. Learn.* 25, 37–56. doi: 10.19173/irrodl.v25i3.7646

Weizenbaum, J. (1966). ELIZA–a computer program for the study of natural language communication between man and machine. *Commun. ACM* 9, 36–45. doi: 10.1145/365153.365168

Wiboolyasar, K., and Jinowat, N. (2020). Learners' oral corrective feedback perceptions and preferences in Thai as a foreign language tertiary setting. *J. Lang. Ling. Stud.* 16, 912–929. doi: 10.17263/jlls.759344

Xu, M., and Zeng, S. (2023). Optimal timing of treatment for errors in second language learning - a systematic review of corrective feedback timing. *Front. Psychol.* 14:1026174. doi: 10.3389/fpsyg.2023.1026174

Yorozu, M. (2001). Interaction with native speakers of japanese: what learners say. *Jpn. Stud.* 21, 199–213. doi: 10.1080/10371390120074363

Zhang, A., Gao, Y., Suraworachet, W., Nazaretsky, T., and Cukurova, M. (2025). *Evaluating Trust in AI, Human, and Co-produced Feedback Among Undergraduate Students*. Available online at: https://arxiv.org/abs/2504.10961 (Accessed December 25, 2025).

Zogaj, A., Mähner, P. M., Yang, L., and Tscheulin, D. K. (2023). It's a match! The effects of chatbot anthropomorphization and chatbot gender on consumer behavior. *J. Business Res.* 155:113412. doi: 10.1016/j.jbusres.2022.113412