

Evaluating ChatGPT's Reliability in Second Language Acquisition (SLA): Insights on Language Skills and Technology's Role

Albatool Ahmad Alhazmi and Muneera Muftah

Department of English, College of Languages and Translation, Najran University, Najran, Saudi Arabia

Aaalhazmi@nu.edu.sa

munmef5@gmail.com (Corresponding Author)

<https://doi.org/10.34190/ejel.23.1.3762>

An open access article under [CC Attribution 4.0](https://creativecommons.org/licenses/by/4.0/)

Abstract: Artificial Intelligence (AI) has become a transformative tool across various fields, including education, offering innovative opportunities for second language (L2) learning and teaching. The potential of AI-driven tools like ChatGPT in Second Language Acquisition (SLA) remains an area of emerging interest, with implications for enhancing pedagogical approaches and providing scalable, interactive support for learners and educators. This study evaluates the reliability and quality of ChatGPT-4.0's responses to 48 key questions related to SLA, encompassing themes such as technology's role and language skills (reading, writing, listening, and speaking). The questions were rated by five expert linguists using a Likert scale to assess accuracy, and relevance. Statistical analyses, including the Kruskal-Wallis H test ($p < 0.05$) and post-hoc pairwise comparisons, revealed significant differences among evaluators ($p < 0.001$), with most responses rated as good (26.25%) or very good (55.41%). Despite these positive outcomes, Fleiss's Kappa coefficient (0.006) highlighted low inter-rater agreement, signaling a need for standardized evaluation frameworks. While ChatGPT shows promise in addressing diverse SLA topics, its limitations—such as occasional inaccuracies—underscore the necessity for ongoing research to optimize its utility in educational contexts. This study contributes to SLA pedagogy by demonstrating the potential and boundaries of integrating AI into language learning, advocating for cautious and evidence-based adoption.

Keywords: Artificial intelligence (AI), ChatGPT, Language education, Language learning technologies, Reliability, Second language acquisition (SLA)

1. Introduction

Artificial intelligence (AI) language models have been transforming the landscape of education and scientific research. These novel models attract the attention of researchers, educators, and users across various fields due to their impressive and versatile capabilities, including content generation in multiple domains such as texts, images, and videos (Liu et al., 2023; Ray, 2023). Among these models, ChatGPT, based on the Generative Pre-trained Transformer (GPT) architecture, has gained significant recognition and acceptance in professional and educational sectors, including medicine, healthcare, finance, business, management, science, translation, and education (Baskara and Mukarto, 2023; Kalla, Smith and Carolina, 2023; Kung et al., 2023; Lund and Wang, 2023; Ray, 2023).

This revolutionary technology is designed to perform several natural language processing (NLP) tasks that mimic human capabilities, such as learning, reasoning, data analysis, decision-making, question-answering, and problem-solving (Kalla, Smith and Carolina, 2023; Kung et al., 2023; Liu et al., 2023; Muftah, 2024). As these models continue to evolve, their integration into daily life and education becomes increasingly inevitable, with notable implications for individualized learning and interaction. However, concerns regarding the accuracy and quality of responses generated by ChatGPT have prompted scholars to call for more rigorous evaluations of its potential benefits, limitations, and risks in academic contexts.

Recent studies have highlighted the need to explore the role of AI technologies in education. For instance, Huallpa et al. (2023) and Baskara and Mukarto (2023) emphasize the importance of investigating ChatGPT's impact on higher education language learning. Similarly, Ray (2023) underscores the necessity of examining ChatGPT's potential to advance deep and personalized learning and scientific research. While the technology offers significant promise, researchers such as Chaka (2023), Liu et al. (2023), and Sallam (2023) advocate for more research to assess the validity of its responses, identify its limitations, and establish strategies to mitigate risks associated with its use.

Despite these discussions, one critical gap in the literature is the lack of empirical studies assessing the reliability of ChatGPT's responses specifically within Second Language Acquisition (SLA) contexts. Current research has largely focused on AI's general application in education, with limited attention to its specific impact on SLA-

related areas such as reading, writing, listening, and speaking skills. Additionally, there is insufficient analysis of how well AI-generated content aligns with established SLA theories and pedagogical frameworks.

To bridge this gap, this study critically evaluates the accuracy and reliability of ChatGPT's responses to SLA-related inquiries, particularly in the context of language skills learning. While AI has demonstrated potential in language education, its effectiveness remains uncertain due to inconsistencies in generated responses, potential misinformation, and lack of alignment with SLA principles. Understanding these factors is essential for assessing whether ChatGPT can serve as a reliable tool for learners and educators.

Given these considerations, the present study seeks to evaluate the reliability of ChatGPT in providing accurate and high-quality responses to inquiries concerning key aspects of SLA. These aspects include the role of technology, reading, writing, listening, and speaking, with a focus on the model's capacity to contribute meaningfully to SLA-related issues. This evaluation addresses a notable gap in research, as limited studies have investigated the reliability and accuracy of AI models within the branches and sub-branches of linguistics.

To achieve this aim, the study addresses the following research questions:

- How reliable are the responses generated by ChatGPT in addressing key questions in SLA?
- To what extent does ChatGPT provide accurate and meaningful information on specific language skills (reading, writing, listening, and speaking) in SLA?
- What are the limitations and areas for improvement in ChatGPT's performance as evaluated by experts in SLA?

By addressing these questions, this study aims to contribute to the growing body of knowledge on the use of AI in SLA and provide insights into its potential applications and limitations in language education.

2. Literature Review

2.1 Advancements in AI and ChatGPT Applications

AI, particularly through language models like the Generative Pretrained Transformer (GPT) series, has revolutionized natural language processing (NLP). These models perform tasks such as contextual understanding, real-time interaction, and personalized support Brown et al. (2020) and Liu et al. (2023) discuss how these advancements enhance AI's adaptability in various fields, particularly education. ChatGPT, a generative chatbot, exemplifies these capabilities, offering human-like conversational responses and demonstrating human-level performance in various professional and academic domains. For instance, Firat (2023) highlights ChatGPT's impact on academic writing and research assistance, while Ray (2023) evaluates its performance in subject-specific tasks.

AI applications, including ChatGPT, are extensively employed across industries such as medicine, business, education, and scientific research (Baskara and Mukarto, 2023; Kalla, Smith and Carolina, 2023; Kung et al., 2023). For instance, ChatGPT aids medical diagnostics by providing interactive and timely responses for disease prevention and treatment planning (Branum and Schiavenato, 2023; Sarraju et al., 2023; Tanaka et al., 2023). Similarly, in education, AI tools like ChatGPT support diverse academic functions, including content creation, metadata generation, and personalized tutoring (Huallpa et al., 2023; Lund and Wang, 2023).

2.2 AI and Education: Opportunities and Challenges

AI's integration into education has been transformative, particularly in enhancing individualized learning experience (Srinivasa, Kurni and Saritha, 2022). Lund and Wang (2023) note that AI enhances self-directed learning by providing real-time feedback, an observation echoed by Firat (2023), who highlights ChatGPT's potential in fostering learner autonomy, motivation, and self-directed learning by providing personalized guidance and prompt feedback. Similarly, Sharma and Yadav (2023) discuss AI's contribution to open and online education, enabling interactive and adaptive learning.

Despite these advantages, AI adoption raises concerns about accuracy, ethical implications, and privacy. For example, Baskara and Mukarto (2023) warn about biases in ChatGPT's responses, which can lead to culturally insensitive outputs. Furthermore, Huallpa et al. (2023) argue that AI tools must be adapted to diverse learners' skills to avoid inequities in learning outcomes. These challenges highlight the need for responsible integration of AI in education.

2.3 ChatGPT in Language Education

The application of ChatGPT in language education holds immense potential, supporting language learning through automated tutoring, writing assistance, language assessment, and personalized language practice in simulated real-life conversational contexts (Barrot, 2023; Brown et al., 2020; Huang, Hew and Fryer, 2022). For example, ChatGPT enhances English proficiency by enriching vocabulary, improving writing skills, fostering reading comprehension, and supporting translation activities (Baskara and Mukarto, 2023). It also aids in improving speaking skills through activities such as storytelling and narration.

Baskara and Mukarto (2023) emphasize the wealth of opportunities ChatGPT provides for language learning, highlighting its role in enhancing communicative proficiency and fostering personalized, collaborative language learning experiences. They illustrate how ChatGPT supports English teaching and learning both inside and outside classrooms, enhancing learners' abilities in writing, grammar, and vocabulary acquisition while developing their social and interactional skills. Similarly, Karataş et al. (2024) found that ChatGPT positively affects students' learning experiences, helping them improve their writing abilities, grammar, and vocabulary while fostering critical social skills.

In addition to its pedagogical applications, ChatGPT offers several technological affordances, including timeliness, ease of use, and personalization. Huang, Hew and Fryer (2022) outline five pedagogical roles of AI chatbots in language learning: as interlocutors, simulations, transmission tools, helplines, and recommendation agents. These roles contribute significantly to enhancing students' social skills through affective and coherent communication. Beyond language learning, AI applications, including ChatGPT, support various areas in applied linguistics, such as language translation, language acquisition, and computer-assisted learning (Chaka, 2023).

Despite these advantages, challenges such as cognitive overload and technological limitations must be addressed to optimize the use of AI technologies in language learning. Researchers, including Karataş et al. (2024), advocate for a balanced approach to integrating AI tools like ChatGPT into language curricula, ensuring their effective application while mitigating potential drawbacks.

2.4 Reliability and Challenges of AI Applications

The reliability of AI tools like ChatGPT has been a focus of extensive research across various disciplines, particularly in education and healthcare. While ChatGPT demonstrates considerable potential, it has also revealed limitations that highlight the need for cautious integration into human-centered fields. For instance, in the medical domain, studies have examined ChatGPT's ability to perform diagnostic and educational tasks. Kung et al. (2023) evaluated ChatGPT's performance in the United States Medical Licensing Examination (USMLE), finding that it scored near the passing threshold without expert training, demonstrating its potential role in medical education and clinical decision-making. Similarly, Tanaka et al. (2023) found that ChatGPT provided effective answers on orthodontic topics such as clear aligners, temporary anchorage devices, and digital imaging. However, deficiencies were noted in generating verifiable responses, as Branum and Schiavenato (2023) reported instances of fabricated and self-generated information, underscoring concerns about its trustworthiness.

In educational settings, ChatGPT raises similar reliability issues. While it offers numerous benefits, such as personalized learning experiences and enhanced accessibility (Barrot, 2023; Lund and Wang, 2023), its inability to consistently provide precise, culturally aware, or ethically sound responses presents challenges. Concerns about accuracy, privacy, and data security are common among educators and students alike (Huallpa et al., 2023). For example, Baskara and Mukarto (2023) highlight that ChatGPT may generate inappropriate or misleading content due to biases in its training data, which often lacks human nuance, cultural awareness, and higher-order thinking. Additionally, Huallpa et al. (2023) emphasize that adapting AI-generated materials to suit diverse learner backgrounds, skills, and interests remains a critical issue.

Ethical considerations further complicate the integration of ChatGPT and similar AI tools in education. Researchers like Lund and Wang (2023) discuss concerns such as transparency, accountability, and intellectual property, emphasizing the importance of addressing these issues to avoid inaccurate or inequitable outcomes. Additionally, academic integrity and the preservation of students' critical and creative thinking skills are significant factors in evaluating the role of AI in learning and teaching (Barrot, 2023; Huallpa et al., 2023). The use of AI in language learning illustrates both the promise and the pitfalls of these technologies. Huang, Hew and Fryer (2022) identify limitations such as cognitive load, technological constraints, and the novelty effect, while Karataş et al. (2024) stress the importance of adopting a balanced approach when integrating AI tools into language curricula. Researchers agree that while AI can enhance learners' experiences by improving skills such

as writing, grammar, and vocabulary acquisition, its limitations necessitate further research and methodological refinement to ensure effective and ethical application (Huang, Hew and Fryer, 2022; Baskara and Mukarto, 2023; Sallam, 2023).

In brief, while ChatGPT and similar AI tools demonstrate substantial potential in education and other fields, their limitations in reliability, ethical considerations, and adaptability demand a cautious and balanced approach. Further research is essential to address these challenges and to develop strategies that maximize their benefits while minimizing risks.

2.5 ChatGPT and SLA: Bridging AI and Language Acquisition

Second Language Acquisition (SLA) is a multidisciplinary field dedicated to understanding the processes and factors involved in learning languages other than one's first language (Ellis, 2015). With its foundations in linguistics, psychology, cognitive science, sociology, and education, SLA research has provided invaluable insights into cognitive processes, the similarities and differences between first language (L1) and second language (L2) acquisition, and the development of effective language teaching methodologies (Muftah, 2023a, 2023b; VanPatten, Smith and Benati, 2020). These findings not only inform language education but also contribute to broader fields such as global communication and cognitive science.

In recent years, advancements in artificial intelligence (AI) have introduced novel tools, such as ChatGPT, with potential applications in SLA. ChatGPT offers opportunities to simulate conversational interactions, provide linguistic feedback, and adapt to learners' individual needs, thereby enhancing learning experiences. Despite these promising features, the reliability of ChatGPT's responses to SLA-specific queries remains underexplored. Current research lacks sufficient evaluation of its accuracy and applicability within the nuanced context of SLA.

This study seeks to bridge this gap by evaluating the reliability and potential of ChatGPT in addressing key aspects of SLA. Specifically, it examines the accuracy and quality of ChatGPT's responses through content analysis of self-generated and self-answered queries. By focusing on critical facets of SLA, this research aims to assess whether ChatGPT can effectively support learning processes, address the complex dynamics of L2 acquisition, and complement traditional teaching methodologies. Building on decades of SLA research, which has uncovered the cognitive, social, and pedagogical factors influencing language learning, this study emphasizes the importance of integrating AI into SLA responsibly. While ChatGPT offers innovative possibilities, understanding its limitations and potential applications is crucial to leveraging AI effectively in language education.

3. Materials and Methods

The study aimed to assess the accuracy and relevance of ChatGPT's responses to key questions related to SLA. To achieve this, ChatGPT was used as an interactive tool to generate responses to 48 pre-designed questions covering various SLA topics, such as the role of technology, and language skills including reading, writing, listening, and speaking. These questions were crafted to evaluate ChatGPT's ability to provide accurate, detailed, and coherent answers. The chatbot responded autonomously to each query without additional contextual prompts, ensuring that its responses were generated based solely on the questions asked.

3.1 Selection of Questions

For the evaluation process, ChatGPT-4 was prompted with 48 questions covering six key areas in SLA: (1) key questions in SLA, (2) the role of technology in SLA, (3) reading skills, (4) writing skills, (5) listening skills, and (6) speaking skills. These questions were selected based on their frequency in academic inquiries and their significance in SLA research, drawing from textbooks, scholarly articles, and online discussions. A timeline-based approach was used to ensure a systematic selection process. Over three months, a comprehensive review of SLA literature and academic discourse was conducted to identify the most frequently discussed topics. Eight representative questions were formulated for each category, ensuring broad coverage of SLA concepts. While specific websites were not used as the sole source, the design was guided by theoretical constructs, research studies, and expert consultation. The final selection was refined and validated by experts before being used to prompt ChatGPT-4. The AI-generated responses to these questions were collected and documented in Table 1.

3.2 Evaluation Process

To assess the quality of the responses, five expert researchers in SLA, each possessing over ten years of academic experience in Applied Linguistics and SLA studies, independently rated the answers on a five-point Likert scale. The scale ranged from 1 (very poor) to 5 (very good), with higher scores indicating superior information quality. The Likert scale used in the evaluation assessed ChatGPT's responses across two dimensions: accuracy and

relevance, each rated on a five-point scale. Accuracy measured the correctness and alignment of the responses with established SLA theories, where 1 indicated significant errors and 5 represented error-free and precise information. Relevance assessed the focus and appropriateness of the responses to the questions posed, ranging from 1 for off-topic or tangential answers to 5 for highly relevant and contextually appropriate responses. These criteria ensured a systematic evaluation process aligned with the study's objectives and theoretical underpinnings, offering a structured approach to assess ChatGPT's performance comprehensively. The evaluation considered the amalgamation of the best available linguistic evidence and academic expertise.

3.3 Addressing Inter-Rater Reliability

Before commencing the scoring process, a calibration session was held to establish a shared understanding of the scoring system among the evaluators, reducing subjectivity and enhancing consistency. The study employed a crowd score strategy, as the evaluation of ChatGPT's responses lacked an established "ground truth" and was inherently subjective. Emphasis was placed on the median scores assigned by the evaluators for each answer. To further address inter-rater reliability concerns, Fleiss's kappa coefficient (Fleiss, 1971) was computed to determine the level of agreement among evaluators, with results interpreted based on Landis and Koch's (1977) guidelines, categorizing agreement levels as slight, fair, moderate, substantial, or almost perfect. Despite some variability in ratings, this method ensured a systematic and evidence-based approach to evaluating ChatGPT's responses.

3.4 Data Analysis Procedure

The scores provided by the evaluators were recorded in Microsoft Excel and analyzed using the Statistical Package for Social Sciences v. 25 (SPSS) software. Each question's median, interquartile range (IQR), and full score range were calculated. Evaluators were assigned random identifiers, and Fleiss's kappa was used to assess the consistency of scores among them (Fleiss, 1971). Moreover, the reliability of the questionnaire, which included the questions, was measured using Cronbach's alpha. The choice of statistical tests was based on the nature of the collected data. The Kruskal-Wallis H test was used to identify any differences in scores among the evaluators due to its suitability for non-parametric data and independent samples (Corder and Foreman, 2011; Kruskal and Wallis, 1952). All statistical analyses were conducted with a significance level of $p < 0.05$. Additionally, when performing post-hoc pairwise comparisons, the Bonferroni correction was applied to manage multiple testing and reduce the likelihood of Type I errors. This structured methodology ensures that the study maintains high levels of validity and reliability, providing a robust assessment of ChatGPT's performance in answering SLA-related questions.

4. Results

Table 1 presents the questions and answers generated by ChatGPT, alongside the median, interquartile range, and overall range of scores provided by evaluators.

Table 1: The queries and answers generated by ChatGPT, along with a summary of evaluators scores

Topics	N	Questions generated by ChatGPT 4.0	Answers generated by ChatGPT 4.0	Median	IQR	Range
Key questions in SLA	1	How did the field of second language acquisition (SLA) research originate?	The field of second language acquisition (SLA) research originated as scholars began to systematically investigate the processes and factors involved in individuals acquiring a second language. This development gained momentum in the mid-20th century, with linguists and psychologists exploring various aspects of language learning and its underlying mechanisms.	4.0	2.0	3.0-5.0
	2	Describe the stages and progression involved in the development of SLA?	The development of SLA involves distinct stages and progression, beginning with the initial exposure to second language (L2). Initial exposure leads to the establishment of rudimentary linguistic competence, with subsequent stages marked by syntactic and morphological refinement, expanding lexicon, and pragmatic development, followed by stages of comprehension, production, and eventual mastery. These stages are shaped by factors such as age,	4.0	2.0	3.0-5.0

Topics	N	Questions generated by ChatGPT 4.0	Answers generated by ChatGPT 4.0	Median	IQR	Range
			exposure, and individual differences, contributing to the complexity of the SLA process.			
	3	What functions do input and output serve in the process of SLA?	Input and output play crucial roles in the process of SLA. Input, as articulated by Stephen Krashen's Input Hypothesis, in the form of exposure to L2, provides the linguistic material for learners, while output posited by Merrill Swain's Output Hypothesis, or the production of language, allows learners to practice and internalize linguistic structures. Both input and output contribute to the gradual development of language proficiency.	5.0	1.0	4.0-5.0
	4	What characterizes the initial state in SLA?	The initial state in SLA is characterized by a learner's starting point in acquiring a second language. Factors such as the learner's prior linguistic knowledge, cognitive abilities, and exposure to L2 contribute to shaping this initial state.	3.0	1.5	3.0-4.5
	5	Is it possible for second language (L2) learners to attain native-like proficiency?	While achieving native-like proficiency is challenging, research suggests that second language (L2) learners can attain a high level of proficiency. However, various factors, including age of acquisition, intensity of exposure, and individual differences, influence the extent to which learners can approach native-like competence.	5.0	1.0	4.0-5.0
	6	How does instructional intervention impact SLA?	Instructional intervention can significantly impact SLA. Effective teaching methods, feedback, and exposure to meaningful language input can enhance the learning process and accelerate language acquisition. However, the effectiveness of instructional methodologies is contingent upon the alignment of pedagogical strategies with learner variables, including motivation, cognitive styles, and prior linguistic experience.	4.0	1.0	3.5-4.5
	7	What significance does explicit learning have in SLA?	Explicit learning, involving conscious awareness of linguistic rules and structures, is recognized as a supplementary mechanism in SLA. While implicit learning remains primary, explicit instruction can expedite the acquisition of intricate grammatical features and foster metalinguistic awareness.	4.0	1.5	3.5-5.0
	8	What are individual differences, and how do they influence SLA?	Individual differences, encompassing variables such as cognitive aptitude, motivation, and learning strategies, exert discernible effects on the trajectory of SLA. These factors interplay with instructional contexts, shaping the pace and extent of linguistic development among L2 learners. Understanding and accommodating these individual differences are imperative for tailored pedagogical approaches	5.0	1.0	4.0-5.0
	1	How has technology impacted SLA in recent years?	Technology has revolutionized second language acquisition by providing interactive and immersive learning experiences. Online resources, language learning apps, and virtual communication tools offer learners access to authentic language use and diverse cultural content.	5.0	1.0	4.0-5.0

Topics	N	Questions generated by ChatGPT 4.0	Answers generated by ChatGPT 4.0	Median	IQR	Range
The role of technology in SLA	2	What are the advantages of incorporating language learning apps in SLA?	Language learning apps provide flexibility, allowing learners to practice at their own pace. They often incorporate gamification elements, making the learning process engaging. Apps also offer instant feedback and personalized learning experiences.	5.0	1.0	4.0-5.0
	3	How can virtual language exchanges enhance SLA?	Virtual language exchanges enable learners to practice with native speakers or peers from around the world through online platforms. This fosters real-life communication, cultural exchange, and the application of language skills in authentic contexts	4.0	1.0	4.0-5.0
	4	In what ways can multimedia content contribute to SLA?	Multimedia content, such as videos, podcasts, and interactive simulations, offers authentic language input with contextual cues. This aids in the development of listening and comprehension skills while exposing learners to diverse accents and language registers	5.0	1.5	3.5-5.0
	5	What role do online language communities play in supporting SLA?	Online language communities provide a platform for learners to connect, share experiences, and practice the target language. These communities facilitate collaborative learning, cultural exchange, and the development of a supportive language-learning network.	5.0	0.5	4.5-5.0
	6	How can virtual reality (VR) technology benefit second language learners?	VR technology creates immersive language learning environments, allowing learners to virtually visit and interact in target language settings. This enhances cultural understanding, situational language use, and overall language proficiency through realistic simulations.	5.0	1.0	4.0-5.0
	7	What considerations should educators keep in mind when integrating technology into language classrooms?	Educators should consider the accessibility of technology, learner preferences, and the alignment of digital tools with pedagogical objectives. Additionally, ensuring a balance between technology and traditional teaching methods is essential for an effective language learning environment.	5.0	1.5	3.5-5.0
	8	Are there any potential challenges associated with technology use in SLA?	Challenges may include issues related to digital access, unequal technology resources, and potential distractions. Additionally, maintaining a balance between technology and face-to-face interaction, and ensuring the quality of online content, are considerations for educators and learners alike.	4.0	1.0	4.0-5.0
Reading in SLA	1	How does the process of reading contribute to SLA, and what specific skills are involved in L2 reading development?	The process of reading contributes significantly to SLA by exposing learners to authentic language use, expanding vocabulary, and enhancing overall language proficiency. Specific skills involved in L2 reading development include decoding, comprehension, inference, and the ability to analyze and critically evaluate texts.	4.0	1.0	3.5-4.5
	2	What role do phonics and decoding skills play in the early stages of L2 reading acquisition?	Phonics and decoding skills play a crucial role in the early stages of L2 reading acquisition. These skills help learners recognize and decipher unfamiliar words, facilitating the development of a strong	5.0	1.5	3.5-5.0

Topics	N	Questions generated by ChatGPT 4.0	Answers generated by ChatGPT 4.0	Median	IQR	Range
			foundation for understanding written language.			
	3	How can educators tailor reading instruction to address the diverse needs and proficiency levels of L2 learners?	Educators can tailor reading instruction for L2 learners by differentiating content based on proficiency levels, providing culturally relevant texts, offering additional support for vocabulary development, and incorporating varied instructional approaches, such as collaborative reading activities and language-rich discussions.	4.0	2.0	3.0-5.0
	4	What impact does extensive reading, such as independent reading of various texts, have on vocabulary acquisition and overall language proficiency in L2 learners?	Extensive reading, characterized by independent reading of a wide range of texts, has a positive impact on vocabulary acquisition and overall language proficiency in L2 learners. Exposure to diverse materials helps learners encounter new words in context and reinforces language patterns.	4.0	1.0	4.0-5.0
	5	How do different reading comprehension strategies, such as summarization and inferencing, influence the development of higher-order thinking skills in SLA?	Different reading comprehension strategies, such as summarization and inferencing, contribute to the development of higher-order thinking skills in SLA. These strategies encourage learners to engage deeply with the text, fostering critical analysis and interpretation.	4.0	2.5	2.0-4.5
	6	What challenges do L2 learners commonly encounter in the process of reading, and how can educators mitigate these challenges to promote successful language acquisition?	L2 learners commonly face challenges in reading, including vocabulary gaps, cultural unfamiliarity, and difficulty with complex syntax. Educators can mitigate these challenges by providing explicit vocabulary instruction, incorporating cultural awareness activities, and scaffolding reading tasks to support comprehension.	5.0	1.0	4.0-5.0
	7	How does the integration of authentic and culturally relevant texts enhance the effectiveness of reading instruction in SLA?	The integration of authentic and culturally relevant texts enhances the effectiveness of reading instruction in SLA by making the content more engaging and meaningful for learners. Exposure to real-world language use and cultural contexts aids in comprehension and language internalization.	5.0	0.5	4.5-5.0
	8	In what ways can technology, such as digital platforms and online resources, be leveraged to support and supplement L2 reading instruction for diverse learner populations?	Technology, such as digital platforms and online resources, can be leveraged to support and supplement L2 reading instruction for diverse learner populations. Interactive e-books, language learning apps, and online reading materials offer opportunities for independent practice, immediate feedback, and personalized learning experiences.	5.0	1.5	3.5-5.0
Writing in SLA	1	How does the process of writing contribute to SLA, and what specific skills are involved in the development of writing proficiency in L2?	The process of writing significantly contributes to SLA by providing learners with opportunities to practice and apply language skills. Specific skills involved in the development of writing proficiency in L2 include grammar usage, vocabulary selection, sentence structure, and organizational coherence.	4.0	2.0	3.0-5.0
	2	What role does grammar and syntax play in the early	In the early stages of L2 writing acquisition, grammar and syntax play a foundational role in establishing basic communication	4.0	2.0	3.0-5.0

Topics	N	Questions generated by ChatGPT 4.0	Answers generated by ChatGPT 4.0	Median	IQR	Range
		stages of L2 writing acquisition, and how do these skills evolve over time?	skills. These skills evolve over time as learners progress, moving from simple sentence constructions to more complex structures, enabling them to express more nuanced ideas and arguments			
	3	How can educators effectively teach and assess writing skills to address the diverse needs and proficiency levels of L2 learners?	Educators can effectively teach and assess writing skills in L2 learners by implementing differentiated instruction that considers diverse proficiency levels. Strategies may include providing clear writing prompts, offering targeted feedback, and incorporating peer review processes to address individual needs.	5.0	0.5	4.5-5.0
	4	What impact does extensive writing practice, including essay composition and creative writing, have on vocabulary acquisition and overall language proficiency in L2 learners?	Extensive writing practice, including essay composition and creative writing, has a positive impact on vocabulary acquisition and overall language proficiency in L2 learners. Engaging in varied writing tasks allows learners to experiment with language, expand their lexicon, and apply grammar rules in context.	5.0	1.5	3.5-5.0
	5	How do different writing strategies, such as pre-writing planning and revision techniques, influence the development of higher-order thinking skills in SLA?	Different writing strategies, such as pre-writing planning and revision techniques, play a crucial role in the development of higher-order thinking skills in SLA. These strategies encourage learners to organize their thoughts, analyze content critically, and refine their written expression	5.0	1.0	4.0-5.0
	6	What challenges do L2 learners commonly encounter in the process of writing, and how can educators provide targeted support to address these challenges?	L2 learners commonly encounter challenges in writing, including difficulty with grammatical accuracy, vocabulary limitations, and struggles with text coherence. Educators can provide targeted support through explicit instruction, individualized feedback, and scaffolded writing tasks	4.0	1.0	4.0-5.0
	7	How does the incorporation of authentic and culturally relevant writing tasks enhance the effectiveness of writing instruction in SLA?	The incorporation of authentic and culturally relevant writing tasks enhances the effectiveness of writing instruction in SLA by making the content more relatable and engaging. Authentic tasks encourage learners to connect language use with real-world contexts, promoting meaningful expression	4.0	1.5	3.5-5.0
	8	In what ways can technology, such as word processing tools and collaborative online platforms, be leveraged to support and supplement L2 writing instruction for diverse learner populations?	Technology, such as word processing tools and collaborative online platforms, can be leveraged to support and supplement L2 writing instruction for diverse learner populations. These tools provide opportunities for collaborative writing, immediate feedback, and the integration of multimedia elements to enhance language learning experiences	3.0	3.0	2.0-5.0
	1	How does the development of listening skills contribute to overall SLA, and what specific aspects of listening proficiency are critical for language learners?	The development of listening skills significantly contributes to overall SLA by enabling learners to comprehend spoken language, recognize linguistic patterns, and enhance communicative competence. Critical aspects of listening proficiency for language learners include accurate comprehension of spoken content, recognition of diverse accents and dialects,	5.0	1.0	4.0-5.0

Topics	N	Questions generated by ChatGPT 4.0	Answers generated by ChatGPT 4.0	Median	IQR	Range
Listening in SLA			and the ability to infer meaning from context.			
	2	What role does exposure to authentic spoken language play in the early stages of L2 listening acquisition, and how does it impact learners at different proficiency levels?	Exposure to authentic spoken language plays a crucial role in the early stages of L2 listening acquisition. It exposes learners to natural language use, intonation, and pronunciation, fostering a deeper understanding of spoken communication. The impact of exposure varies at different proficiency levels, influencing beginners in basic comprehension and more advanced learners in nuanced understanding and interpretation.	5.0	1.5	3.5-5.0
	3	How can educators design listening activities that cater to the diverse needs and proficiency levels of L2 learners?	Educators can design listening activities for L2 learners that cater to diverse needs and proficiency levels by incorporating a range of difficulty levels, diverse accents, and varied topics. Tailoring activities to individual learning styles and providing opportunities for repetition and reinforcement help address the unique challenges learners may face.	5.0	1.0	4.0-5.0
	4	What influence does active engagement with various audio materials, such as podcasts or language learning apps, have on vocabulary acquisition and comprehension in L2 listening?	Active engagement with various audio materials, such as podcasts or language learning apps, positively influences vocabulary acquisition and comprehension in L2 listening. Exposure to real-world language use in context enhances vocabulary retention and allows learners to apply their knowledge in practical situations.	5.0	1.5	3.5-5.0
	5	How do different listening comprehension strategies, such as prediction and summarization, contribute to the development of higher-order thinking skills in SLA?	Different listening comprehension strategies, such as prediction and summarization, contribute to the development of higher-order thinking skills in SLA. These strategies encourage learners to engage critically with spoken content, enhancing their ability to analyze, synthesize, and evaluate information.	5.0	1.5	3.5-5.0
	6	What common challenges do L2 learners encounter in the process of listening, and how can instructional approaches be adapted to address these challenges effectively?	Common challenges encountered by L2 learners in the process of listening include difficulty with rapid speech, unfamiliar accents, and processing complex linguistic structures. Instructional approaches can be adapted by providing pre-listening activities, offering explicit strategies for comprehension, and gradually increasing the complexity of listening tasks.	5.0	0.5	4.5-5.0
	7	How does the integration of culturally relevant and authentic listening materials enhance the effectiveness of instruction in SLA?	The integration of culturally relevant and authentic listening materials enhances the effectiveness of instruction in SLA by making the content more engaging and meaningful. Learners connect language use with real-world contexts, fostering a deeper appreciation of cultural nuances and communicative purposes.	4.0	1.5	3.5-5.0
	8	In what ways can technology, such as digital platforms and online resources, be utilized to support and augment L2 listening instruction for diverse learner populations?	Technology, such as digital platforms and online resources, can be utilized to support and augment L2 listening instruction for diverse learner populations. Interactive tools, language learning apps, and online platforms offer opportunities for personalized learning experiences, allowing learners to practice listening skills in varied contexts and at their own pace.	5.0	1.5	3.5-5.0

Topics	N	Questions generated by ChatGPT 4.0	Answers generated by ChatGPT 4.0	Median	IQR	Range
Speaking in SLA	1	How does the development of speaking skills contribute to overall SLA, and what specific aspects of speaking proficiency are critical for language learners	The development of speaking skills significantly contributes to overall SLA by providing learners with the ability to express themselves verbally, engage in effective communication, and participate actively in various social contexts. Critical aspects of speaking proficiency for language learners include pronunciation, fluency, vocabulary usage, and the ability to convey complex ideas.	5.0	1.0	4.0-5.0
	2	What role does pronunciation play in the early stages of L2 speaking acquisition, and how does it impact learners at different proficiency levels?	Pronunciation plays a crucial role in the early stages of L2 speaking acquisition as it influences the clarity of communication. Learners at different proficiency levels may experience challenges ranging from basic sound production to mastering intonation and rhythm, impacting their overall communicative competence.	5.0	2.0	3.0-5.0
	3	How can educators design speaking activities that cater to the diverse needs and proficiency levels of L2 learners?	Educators can design speaking activities for L2 learners that cater to diverse needs and proficiency levels by incorporating varied topics, real-life scenarios, and interactive tasks. Differentiating instruction based on learners' strengths and weaknesses, providing opportunities for peer collaboration, and offering a supportive environment contribute to effective speaking skill development.	5.0	1.5	3.5-5.0
	4	What influence does regular practice and interaction, such as participating in conversations or language exchange, have on vocabulary acquisition and fluency in L2 speaking?	Regular practice and interaction, such as participating in conversations or language exchange, have a positive influence on vocabulary acquisition and fluency in second language speaking. Engaging in authentic communication fosters vocabulary retention, allows learners to experiment with language use, and enhances overall speaking proficiency.	5.0	1.5	3.5-5.0
	5	How do different speaking strategies, such as role-playing and storytelling, contribute to the development of higher-order thinking skills in SLA?	Different speaking strategies, such as role-playing and storytelling, contribute to the development of higher-order thinking skills in SLA. These activities encourage learners to think critically, organize their thoughts coherently, and express complex ideas, promoting cognitive growth.	5.0	1.5	3.5-5.0
	6	What common challenges do L2 learners encounter in the process of speaking, and how can instructional approaches be adapted to address these challenges effectively?	Common challenges faced by L2 learners in the process of speaking include anxiety, hesitation, and difficulty with spontaneous expression. Instructional approaches can be adapted by incorporating supportive activities, providing opportunities for low-pressure practice, and gradually increasing the complexity of speaking tasks.	4.0	1.0	4.0-5.0
	7	How does the integration of culturally relevant and authentic speaking tasks enhance the effectiveness of instruction in SLA?	The integration of culturally relevant and authentic speaking tasks enhances the effectiveness of instruction in second language acquisition by making the content more engaging and meaningful. Learners connect language use with real-world contexts, fostering a deeper appreciation of cultural nuances and communicative purposes.	5.0	1.5	3.5-5.0

Topics	N	Questions generated by ChatGPT 4.0	Answers generated by ChatGPT 4.0	Median	IQR	Range
	8	In what ways can technology, such as language learning apps and collaborative online platforms, be utilized to support and supplement L2 speaking instruction for diverse learner populations?	Technology, such as language learning apps and collaborative online platforms, can be utilized to support and supplement L2 speaking instruction for diverse learner populations. Interactive tools, video conferencing, and language exchange platforms offer opportunities for personalized practice, immediate feedback, and engagement with a global community of language learners.	5.0	1.0	4.0-5.0

Generally, the evaluators rated ChatGPT as providing valuable information on the assessed topics: key questions in SLA= 4.200 ± 0.823, the role of technology in SLA= 4.525 ± 0.640, reading in SLA= 4.250 ± 0.927, writing in SLA= 4.225 ± 0.974, listening in SLA= 4.475 ± 0.751 and speaking in SLA= 4.425 ± 0.813. The total median score for all main topics was 4.7, showing no statistical difference among them ($p > 0.05$) as indicated in Table 2.

Table 2 displays descriptive statistics for various aspects of SLA, including key questions, the role of technology, reading, writing, listening, and speaking. The data for the writing in SLA category showed greater variability in scores.

Table 2: Descriptive statistics regarding SLA related issues

Topics	Mean	Standard deviation	Median	P value
Key questions in SLA	4.200	0.823	4.0	0.210
The role of technology in SLA	4.525	0.640	5.0	
Reading in SLA	4.250	0.927	4.5	
Writing in SLA	4.225	0.974	5.0	
Listening in SLA	4.475	0.751	5.0	
Speaking in SLA	4.425	0.813	5.0	

Statistical difference for $p < 0.05$.

Table 3 displays the distribution of scores assigned by evaluators, ranging from "very poor" to "very good." The overall findings reveal that the highest percentage of scores in the entire dataset fell within the "Very Good" category (55.41%), followed by "Good" (26.25.1%) and "Acceptable" (16.67%). Conversely, the categories of "Poor" and "Very Poor" had the lowest percentages: 1.25% and 0.42%, respectively.

Table3: Total count (n) and proportion (%) of median scores assigned by each evaluator

Evaluators	(1) Very Poor	(2) Poor	(3) Acceptable	(4) Good	(5) Very Good	
1	n	0	0	6	8	34
	(%)	0.0%	0.00%	12.50%	16.67%	70.83%
2	n	0	1	8	20	19
	(%)	0.0%	2.08%	16.67%	41.67%	39.58%
3	n	0	0	0	9	39
	(%)	0.0%	0.0%	0.0%	18.75%	81.25%
4	n	0	0	1	10	37
	(%)	0.0%	0.0%	2.08%	20.83%	77.09%
5	n	1	2	25	16	4
	(%)	2.08%	4.18%	52.08%	33.33%	8.33%
Total	n	1	3	40	63	133
	(%)	0.42%	1.25%	16.67%	26.25%	55.41%

These findings indicate that most evaluators assigned scores within the higher quality categories (Good and Very Good), suggesting a generally positive evaluation of the topics. The low percentages in the "Poor" and "Very Poor" categories suggest that the topics were perceived to be reasonable and satisfactory in quality by the evaluators. These lower ratings primarily stemmed from ChatGPT's oversimplifications, lack of depth in complex SLA concepts, and occasional inaccuracies in theoretical explanations. For example, responses regarding the characterization of the "initial state in SLA" demonstrated a tendency toward broad generalizations, failing to differentiate adequately between competing theoretical perspectives, such as Universal Grammar-based models and emergentist approaches. Additionally, ChatGPT's responses on leveraging technology in writing instruction were noted to lack specificity, often providing generic recommendations rather than engaging with established research on the pedagogical effectiveness of AI-driven writing assistants. Another area of concern was the response to questions about fossilization in SLA, where ChatGPT provided a surface-level definition but omitted key discussions on how various linguistic, cognitive, and sociocultural factors interact to contribute to fossilization processes. These findings highlight specific gaps in ChatGPT's capabilities and underscore the need for its enhancement in handling nuanced and research-intensive queries. The qualitative observations reinforce the statistical results, indicating that while ChatGPT performs well in providing general explanations, it struggles with depth, precision, and the integration of competing theoretical viewpoints.

In terms of the scores given by evaluators, there was a minimal level of agreement observed (Landis and Koch, 1977), with a combined Fleiss's Kappa coefficient of 0.006. This discrepancy reflects diverse interpretations of the evaluation criteria and underscores the subjective nature of assessing AI-generated content. The Kruskal-Wallis test revealed a notable inconsistency in scores among the evaluators ($p < 0.001$), which has been attributed to differences in individual expertise, perspectives, and familiarity with SLA-specific issues. A comprehensive pairwise comparison of the evaluators' scores is provided in Table 4.

Table 4: Pairwise comparison of evaluators scores.

Evaluators	Bonferroni pairwise comparison	
	t	p-value
1vs.2	2.817	0.028
1vs.3	-2.223	0.428
1vs.4	-1.359	0.725
1vs.5	-1.248	0.824
2vs.3	-4.916	0.000
2vs.4	-4.391	0.000
2vs.5	2.158	0.591
3vs.4	0.658	0.990
3vs.5	-1.723	0.482
4vs.5	-1.969	0.539

Statistical difference for $p < 0.05$.

5. Discussion

Research Question 1: How reliable are the responses generated by ChatGPT in addressing key questions in SLA?

This article investigated how ChatGPT, a sophisticated AI, is currently being used, with a focus on how accurately and efficiently it responds to questions about SLA, the impact of technology, and language skills including reading, writing, listening, and speaking in SLA. In this research, the chatbot quickly provided answers to all questions within seconds. These responses were later evaluated by experienced linguists. The present study represents a groundbreaking cross-sectional evaluation of ChatGPT's precision in addressing contemporary linguistic questions.

The data reflects positive evaluation of responses presented by ChatGPT from the experts in SLA participated in this study. The assessment of ChatGPT's comprehension and interaction indicates a vastly accurate and high-quality information provided by this NLP model regarding various domains in SLA. This finding aligns with several studies that have shown that AI exhibited impressive performance, providing accurate responses across various

topics as evaluated by experts from different academic disciplines, albeit with significant constraints (Firat, 2023; Huallpa et al., 2023; Lund and Wang, 2023). While AI is not a recent innovation, ChatGPT has gained widespread popularity and acceptance (Kalla, Smith and Carolina, 2023; Liu et al., 2023; Ray, 2023). As early mentioned in this paper, various forms of AI have been extensively utilized across disciplines such as management, science, biological and medical diagnostics as well as education (Baskara and Mukarto, 2023; Kalla, Smith and Carolina, 2023; Kalla, Smith and Carolina, 2023; Lund and Wang, 2023; Ray, 2023). These AIs can be trained using diverse datasets and applied to a multitude of tasks within different domains (Lund and Wang, 2023; Makaremi, Lacaule and Mohammad-Djafari, 2019; Ray, 2023; Sharma and Yadav, 2023).

However, the reliability of using these human-like applications has been the concern of several studies across various disciplines including language learning fields. One particular study, (Chaka, 2023), concluded that AI chatbots are currently not considered trustworthy or reliable sources of information within the realm of applied English language studies (AELS). Hence, there is a call for serious work for adopting strategies that ensure effective use of such models in language education (for instances Al-khresheh, 2024; Barrot, 2023; Karataş et al., 2024). Various researchers (e.g., Baskara and Mukarto, 2023; Huallpa et al., 2023; Sallam, 2023) also assert the need to conduct more research and develop methods that equip learners with the tools they need for efficient use of these models and enable useful implementing for such new technologies across different educational contexts.

Research Question 2: To what extent does ChatGPT provide accurate and meaningful information on specific language skills (reading, writing, listening, and speaking) in SLA?

It is noteworthy that previous studies did not address the realm of SLA. In this study, the vast majority of responses were deemed highly satisfactory by the five evaluators. The responses that received the lowest scores were “Q: Key questions in SLA: What characterizes the initial state in SLA? / A: The initial state in SLA is characterized by a learner's starting point in acquiring a second language. Factors such as the learner's prior linguistic knowledge, cognitive abilities, and exposure to L2 contribute to shaping this initial state.” (median: 3.0) and “Q: Writing in SLA: In what ways can technology, such as word processing tools and collaborative online platforms, be leveraged to support and supplement L2 writing instruction for diverse learner populations? / A: Technology, such as word processing tools and collaborative online platforms, can be leveraged to support and supplement L2 writing instruction for diverse learner populations. These tools provide opportunities for collaborative writing, immediate feedback, and the integration of multimedia elements to enhance language learning experiences.” (median: 3.0).

As can be noted from the data, responses from ChatGPT are straightforward, and a linguist with poor training and lacking linguistic expertise may mistakenly perceive all answers as accurate, potentially leading to misinformation. Conversely, linguists with proper training can leverage their academic background to expand or adapt existing theories to accommodate new functions and contexts. However, when dealing solely with written text from ChatGPT, the situation becomes more complex and without experts to discern truth from falsehood, we lack the intermediary needed to navigate intricate subjects effectively (Barrot, 2023; Huallpa et al., 2023; Lund and Wang, 2023; Tanaka et.al., 2023). While large linguistic algorithms excel in knowledge-based assessments, they often struggle with language instruction, learning topics, and literature (Barrot, 2023; Baskara and Mukarto, 2023). To enhance their performance, these AI models require training on high-quality datasets (Huallpa et al., 2023; Liu et al., 2023; Sallam, 2023; Ray, 2023). Nonetheless, their current training on possibly biased datasets could account for the inaccuracies observed when responding to specific research-related queries.

Research Question 3: What are the limitations and areas for improvement in ChatGPT's performance as evaluated by experts in SLA?

Hence, considering that AI is not a novel concept, what has propelled this recent advancement into the mainstream? The intriguing idea of a textbot capable of proficiently generating content on a wide range of topics may be attributed to human curiosity and desire to maximize knowledge (Lund & Wang, 2023; Karataş et al., 2024; Ray, 2023; Sun, 2023). Despite ChatGPT providing robust answers across the six examined subjects, it's crucial to note that this AI learns from an extensive text dataset derived from books, articles, and web pages. Linguists and L2 learners seek more precise responses, but ChatGPT incorporates both accurate scientific information and potentially misleading content found in advertisements, social media, and websites (Barrot, 2023; Baskara and Mukarto, 2023; Huallpa et al., 2023).

Still, ChatGPT, an AI chatbot, effectively summarizes information, generates sophisticated text, and presents itself convincingly (Baskara and Mukarto, 2023; Liu et al., 2023; Lund and Wang, 2023). In this study, its linguistic responses demonstrated a general accuracy: among 240 answers, most were categorized as either very good (55.41%) or good (26.25%). Yet, it requires improvement in accuracy regarding issues related to SLA. The present findings were broad, realistic, and comprehensive, showcasing a solid understanding of the subject matter, but without delving into specific details. However, because ChatGPT functions differently from a search engine, the accuracy of its responses is contingent upon the accuracy of the source information (Barrot, 2023; Liu et al., 2023; Tanaka et al., 2023). While ChatGPT generally delivers accurate responses regarding SLA, it encounters limitations such as: (1) its inability to critically evaluate or analyze findings from previous literature, (2) reliance on a knowledge base limited to data up to 2020 with no updates, (3) occasional misinterpretation of various theories, (4) incapacity to discern between reputable and predatory journal sources, and (5) concerns about scientific precision, potential biases, and the distribution of misinformation to users.

At this point, AI should be viewed as a supplementary tool rather than a replacement for linguists. Researchers like Sharma and Yadav, (2023) and Karataş et al. (2024) emphasize the extent to which AI applications contribute to education that is complementing rather than replacing the facilitator role of humans. The utilization of ChatGPT can be seen as unavoidable, yet linguists must actively pursue theoretical and scientific validation. Although there is considerable interest in utilizing ChatGPT, it lacks extensive training on language acquisition data and specific responses. It is probable that second language learners will resort to ChatGPT for guidance in interpreting data and grasping technological and various language learning strategies. While ChatGPT is perceived as a potentially advantageous aid in language learning environments, apprehensions persist regarding its precision, dependability, and educational implications (Al-khresheh, 2024; Barrot, 2023; Baskara and Mukarto, 2023; Huang, Hew and Fryer, 2022; Karataş et al., 2024)

In sum, ChatGPT has demonstrated impressive proficiency in the evaluated language acquisition areas. However, variations in individual evaluations underscore the inherent complexity and subjectivity of any assessment process. It is imperative for L2 learners and linguists to acknowledge the limitations and ethical considerations associated with ChatGPT, and to consistently validate information using reliable sources. Prior to integrating these AI models into the fields of L2 learning and teaching, concerted efforts are necessary to improve their reliability and to enhance teachers' and learners' awareness about ethical and technical issues that must be considered when implementing such NLP models into education (Al-khresheh, 2024; Chaka, 2023; Huang, Hew and Fryer, 2022; Karataş et al., 2024; Sharma and Yadav, 2023).

6. Conclusion and Implications

This study explored ChatGPT's accuracy, relevance, and reliability in addressing foundational and contemporary topics in SLA, demonstrating its potential as an interactive tool in language education. While ChatGPT produced coherent and contextually relevant responses, limitations such as inconsistencies in evaluator ratings and the absence of expert validation during question design highlighted areas for improvement. These findings underscore the importance of robust research methodologies when evaluating AI tools in educational contexts. The study's implications suggest that AI can serve as a valuable supplement in language learning, providing support for complex SLA concepts. However, educators and researchers must approach its application cautiously, ensuring it complements rather than replaces traditional methods. Future research should prioritize expert validation and inter-rater reliability analyses to enhance methodological rigor and better harness AI's pedagogical potential in SLA. Future research could also explore practical applications of AI tools, such as ChatGPT, in classroom settings, curriculum design, and teacher training programs. Additionally, providing detailed guidelines on how AI-generated content can complement traditional SLA methods would offer educators concrete strategies for integration. These steps aim to bridge the gap between theoretical insights and practical applications, making the study's contributions more impactful for the field.

7. Limitations and Recommendations

Despite its contributions, this study has several limitations that should be addressed in future research. One key limitation is the lack of expert validation for the questions used in the study. While the questions were designed to cover foundational aspects of Second Language Acquisition (SLA) and reflect contemporary trends in SLA discourse, expert input could have ensured greater accuracy and relevance, thereby strengthening the study's methodological rigor. Future research should incorporate expert validation in the question design phase to enhance the precision and reliability of assessment tools.

Another limitation is the absence of an inter-rater reliability analysis prior to the evaluation process. Given that the evaluators had different levels of experience with AI-based tools, the consistency of their ratings could have been improved with such an analysis. Although Fleiss's Kappa coefficient revealed minimal agreement among evaluators, this inconsistency highlights a concern that future studies could address. To mitigate evaluator variability, future research should implement pre-evaluation calibration sessions to align evaluators' expectations and criteria, ensuring a more standardized rating process. Additionally, future studies could explore evaluators' familiarity with AI-based tools like ChatGPT, as understanding their prior exposure to such technologies could refine study design and improve the interpretation of rating discrepancies.

Beyond methodological considerations, the study highlights the potential of AI tools like ChatGPT in SLA research and language education. Future research could explore practical applications of AI in classroom settings, curriculum design, and teacher training programs. For instance, AI-powered chatbots could provide real-time feedback to students on writing tasks, while adaptive learning models could tailor language instruction to individual learners' proficiency levels. Additionally, integrating AI-assisted assessment tools into language programs could help educators streamline feedback processes and improve learning outcomes. These avenues for future research will not only expand the current understanding of AI's role in language learning but also provide practical insights into its pedagogical integration.

Competing interests: The authors have no conflict of interest to declare.

Authors Contributions: **Albatool Alhazmi** was responsible for the Introduction, Literature Review, Resources, and Writing—Reviewing and Editing. **Muneera Muftah** contributed to Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Validation, and Writing—Original Draft Preparation and Editing.

Ethics declarations: Ethics statement: The authors obtained ethical approval (Approval No. 202402-076-018543-041930) from Najran University's Research Ethics Committee, ensuring adherence to ethical standards, including informed consent, maintaining confidentiality, and complying with regulations.

Ethics approval and consent to participate: Confidentiality and the participants' identities were protected by collecting data anonymously. All participants were assured that their responses would only be used for scientific research purposes. All the participants have voluntarily taken part in the study.

AI statement: This research did not utilize any AI-generated content in the writing, analysis, or interpretation of the findings. The study was conducted using original human-authored research methods, including manual data collection, statistical analysis, and literature synthesis. While ChatGPT was the subject of investigation in this study, it was not employed in drafting, editing, or formulating the manuscript. All conclusions and discussions presented in this paper are based on empirical findings and human interpretation

References

- Al-Khresheh, M., 2024. Bridging technology and pedagogy from a global lens: Teachers' perspectives on integrating ChatGPT in English language teaching. *Computers and Education: Artificial Intelligence*, 6(1), pp.1-12. <https://doi.org/10.1016/j.caeai.2024.100218>
- Barrot, J.S., 2023. Using ChatGPT for second language writing: Pitfalls and potentials. *Assessing Writing*, 57(1), pp.1-6. <https://doi.org/10.1016/j.asw.2023.100745>
- Baskara, R. and Mukarto, M., 2023. Exploring the implications of ChatGPT for language learning in higher education. *Indonesian Journal of English Language Teaching and Applied Linguistics*, 7(2), pp.343-358. <http://dx.doi.org/10.21093/ijeltal.v7i2.1387>
- Branum, C. and Schiavenato, M., 2023. Can ChatGPT accurately answer a PICOT question? Assessing AI response to a clinical question. *Nurse Educator*, 48(5), pp.231-233. <https://doi.org/10.1097/NNE.0000000000001436>
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I. and Amodei, D., 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, pp.1877-1901.
- Chaka, C., 2023. Generative AI chatbots-ChatGPT versus YouChat versus Chatsonic: Use cases of selected areas of applied English language studies. *International Journal of Learning, Teaching and Educational Research*, 22(6), pp.1-19. <https://doi.org/10.26803/ijlter.22.6.1>
- Corder, G.W. and Foreman, D.I., 2011. *Nonparametric statistics for non-statisticians*. Hoboken: John Wiley & Sons, pp.99-105. <http://dx.doi.org/10.1002/9781118165881>
- Ellis, R., 2015. *Understanding second language acquisition* (2nd ed.). UK: Oxford University Press.
- Firat, M., 2023. How ChatGPT can transform autodidactic experiences and open education?. [online] Available at: <https://doi.org/10.31219/osf.io/9ge8m>

- Fleiss, J.L., 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), pp.378-382. <https://doi.org/10.1037/h0031619>
- Huallpa, J., Arocutipá, J., Panduro, W., Huete, L., Limo, F., Herrera, E., Callacna, R., Flores, V., Romero, M., Quispe, I. and Hernández, F., 2023. Exploring the ethical considerations of using ChatGPT in university education. *Periodicals of Engineering and Natural Sciences*, 11(4), pp.105-115. <http://dx.doi.org/10.21533/pen.v11i4.3770>
- Huang, W., Hew, K. and Fryer, L., 2022. Chatbots for language learning—Are they really useful? A systematic review of chatbot-supported language learning. *Journal of Computer Assisted Learning*, 38(1), pp.237-257. <https://doi.org/10.1590/2177-6709.28.5.e2323183.oar>
- Jung, S.K. and Kim, T.W., 2016. New approach for the diagnosis of extractions with neural network machine learning. *American Journal of Orthodontics and Dentofacial Orthopedics*, 149(1), pp.127-133. <https://doi.org/10.1016/j.ajodo.2015.07.030>
- Kalla, D., Smith, N. and Carolina, N., 2023. Study and analysis of ChatGPT and its impact on different fields of study. *International Journal of Innovative Science and Research Technology*, 8(3), pp.827-833.
- Karataş, F., Abedi, F., Ozek Gunyel, F., Karadeniz, D. and Kuzgun, Y., 2024. Incorporating AI in foreign language education: An investigation into ChatGPT's effect on foreign language learners. *Education and Information Technologies*, pp.1-24. <https://doi.org/10.1007/s10639-024-12574-6>
- Kruskal, W.H. and Wallis, W.A., 1952. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260), pp.583-621. <https://doi.org/10.1080/01621459.1952.10483441>
- Kung, T., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G. and Maningo, J., 2023. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS Digital Health*, 2(2), pp.1-12. <https://doi.org/10.1371/journal.pdig.0000198>
- Landis, J.R. and Koch, G.G., 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1), pp.159-174. <https://doi.org/10.2307/2529310>
- Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z., Wu, Z., Zhao, L., Zhu, D., Lie, X., Qianga, N., Shenf, D., Liu, T. and Ge, B., 2023. Summary of ChatGPT-related research and perspective towards the future of large language models. *Meta-Radiology*, 1(2), pp.1-14. <https://doi.org/10.1016/j.metrad.2023.100017>
- Lund, B. and Wang, T., 2023. Chatting about ChatGPT: How may AI and GPT impact academia and libraries?. *Library Hi Tech News*, 40(3), pp.26-29. <https://doi.org/10.1108/LHTN-01-2023-0009>
- Makaremi, M., Lacaule, C. & Mohammad-Djafari, A., 2019. Deep learning and artificial intelligence for the determination of the cervical vertebra maturation degree from lateral radiography. *Entropy*, 21(12), p.1222. Available at: <https://doi.org/10.3390/e21121222>.
- Muftah, M., 2023a. Error analysis in second language acquisition (SLA): Types and frequencies of grammatical errors of simple present and past tense in the elicited written production task of Arab EFL undergraduate learners. *Colombian Applied Linguistics Journal*, 25(1), pp.42-56. Available at: <https://doi.org/10.14483/22487085.19202>.
- Muftah, M., 2023b. Data-driven learning (DDL) activities: Do they truly promote EFL students' writing skills development? *Education and Information Technologies*, 28(10), pp.13179-13205. Available at: <https://doi.org/10.1007/s10639-023-11620-z>.
- Muftah, M., 2024. Impact of social media on learning English language during the COVID-19 pandemic. *PSU Research Review*, 8(1), pp.211-226. Available at: <https://doi.org/10.1108/PRR-10-2021-0060>.
- OpenAI, 2024. GPT-4. Available at: <https://openai.com/research/gpt-4>.
- Ray, P., 2023. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 3, pp.121-154. Available at: <https://doi.org/10.1016/j.iotcps.2023.04.003>.
- Sallam, M., 2023. The utility of ChatGPT as an example of large language models in healthcare education, research and practice: Systematic review on the future perspectives and potential limitations. *MedRxiv*, pp.1-34. Available at: <https://doi.org/10.1101/2023.02.19.23286155>.
- Sarraj, A., Bruemmer, D., Van Iterson, E., Cho, L., Rodriguez, F. & Laffin, L., 2023. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. *Jama*, 329(10), pp.842-844. Available at: <https://doi:10.1001/jama.2023.1044>.
- Sharma, S. & Yadav, R., 2023. ChatGPT – A technological remedy or challenge for education system. *Global Journal of Enterprise Information System*, 14(4), pp.46-51. Available at: <https://www.gjeis.com/index.php/GJEIS/article/view/698>.
- Srinivasa, K., Kurni, M. & Saritha, K., 2022. Harnessing the power of AI to education. In: *Learning, Teaching, and Assessment Methods for Contemporary Learners*, pp.311-342. Singapore: Springer. Available at: https://doi.org/10.1007/978-981-19-6734-4_13.
- Sun, C., 2023. Curiosity driven learning in artificial intelligence and its applications. Doctoral thesis, Nanyang Technological University, Singapore. Available at: <https://hdl.handle.net/10356/172831>.
- Tanaka, O., Gasparello, G., Hartmann, G., Casagrande, F. & Pithon, M., 2023. Assessing the reliability of ChatGPT: A content analysis of self-generated and self-answered questions on clear aligners, TADs and digital imaging. *Dental Press Journal of Orthodontics*, 28, pp.1-22. Available at: <https://doi.org/10.1590/2177-6709.28.5.e2323183.oar>.
- VanPatten, B., Smith, M. and Benati, A.G., 2020. *Key questions in second language acquisition: An introduction*. UK: Cambridge University Press.