TESOL in Context
Journal of ACTA
Australian Council of TESOL Associations
ISSN 2209-0916

Integrating an AI chatbot to enhance English language speaking skills: A classroom-based study with Georgian university students

Eliza Kintsurashvili[1] [1]Humanities and Social Sciences-International Black Sea University, Tbilisi, Georgia

Ekaterine Pipia[2] [2]Humanities and Social Sciences-International Black Sea University, Tbilisi, Georgia

Malila Prado[3] [3]Department of Languages and Cultures, Beijing Normal-Hong Kong Baptist University, China

Abstract

The recent expansion of Artificial Intelligence (AI) has prompted renewed attention to English language teaching. To explore whether AI tools may assist in enhancing speaking proficiency among university English language students in Georgia, this paper examines a pedagogical intervention that employed an AI chatbot in the classroom. The experiment involved 40 Georgian university students divided into control and experimental groups over a 12-week period. Pre- and post-tests were conducted, complemented by student interviews and instructor observations. Analysis revealed significant improvements in speaking proficiency, including grammatical accuracy, vocabulary development, pronunciation, fluency, and overall confidence. Moreover, the findings indicate that the use of the AI chatbot helped reduce speaking anxiety by providing a low-pressure environment for practice. Students in the experimental group reported increased self-confidence and motivation, attributing their progress to the consistent feedback and interactive learning experience provided by the AI chatbot.

Correspondence
Malila Prado
malilaprado@bnbu.edu.cn

Publication
Received: 30 January 2026
Revision: 07 April 2026
Accepted: 16 April 2026
Published: 08 May 2026

These findings suggest that, when used as classroom-based support, AI chatbots may offer valuable opportunities for additional language practice and learning.

Keywords: *Artificial Intelligence (AI); AI chatbots; English language teaching; pedagogical intervention; speaking proficiency.*

Introduction

In recent years, technology has been progressively integrated into language education, providing resources such as digital platforms and interactive environments to enhance English language learning (Galante et al., 2023). This digital shift has accelerated efforts to promote more opportunities for student-centered classrooms, with increased and facilitated engagement in global communication (Chiu, Moorhouse et al., 2023; Chiu, Xia et al., 2023; Schmid, 2017). In parallel, the new era of Artificial Intelligence (AI) is prompting a re-evaluation of traditional language teaching approaches, particularly those that prioritize grammar and writing often at the expense of speaking development (Chiu, Moorhouse et al., 2023; Chiu, Xia et al., 2023; Du & Daniel, 2024; Zou et al., 2023), in turn causing many students to remain ill-prepared for real-world communicative demands.

Despite increased awareness of the importance of speaking proficiency, traditional language teaching, particularly in large-sized classes, tends to offer limited opportunities for students to enhance their speaking skills. Although AI chatbots (Artiles Rodríguez et al., 2021) have been shown to support learner autonomy in writing development (Barrot, 2023a, 2023b; Tseng & Warschauer, 2023), their use in developing speaking skills remains underexplored. The few exceptions (i.e., Ahmad Al-husban, 2025; Du & Daniel, 2024; Zou et al., 2023) note that the integration of AI chatbots as a tool for English language speaking enhancement represents a change in approach toward more personalized, interactive, participatory models of language learning.

This study investigated the effectiveness of targeted intervention using Microsoft Copilot, an AI chatbot, to enhance the speaking proficiency of B1-level university English language learners in Georgia. A one-semester intervention was delivered to two groups: one received instruction supported by Copilot AI, while the other followed the same instructional model but without the inclusion of an AI chatbot. The objective of the experiment was to assess the impact of AI integration on students' speaking engagement and performance in the English language classroom while also exploring the potential for such tools to improve learning outcomes compared to tools

and approaches previously adopted within the institution and to reduce teacher workload (Ahmad Al-husban, 2025; Artiles Rodríguez et al., 2021; Du & Daniel, 2024; Zou et al., 2023).

In this paper, following a review of relevant studies on speaking proficiency in the language classroom aided by the introduction of AI tools, with a focus on speaking development and learner motivation, we describe the methodological design of our classroom-based intervention and present the results of our comparative analysis. Finally, we discuss the pedagogical implications of integrating AI tools such as Copilot in English Language Teaching (ELT) contexts and highlight directions for future research.

Speaking proficiency in English language teaching

In ELT contexts, speaking should be at the heart (Fan & Yan, 2020; Fernández-García & Fonseca-Mora, 2022; Robayo Acuña, 2025). By speaking, we mean the sociolinguistic, interactional competence that is necessary when participants of a communicative event negotiate meaning through features such as pronunciation (in terms of intelligibility), communicative strategies, fluency enhancement markers, grammar and vocabulary (Fulcher, 2014). These features, in combination, can be analyzed – in speaking – through idea units, which are "units [...] usually spoken with a coherent intonation contour, and they are often limited on both sides by pauses or hesitation markers." (Luoma, 2004, p. 12). In the teaching of English language speaking, although recent textbooks may include more fluency- and interaction-oriented tasks – however unauthentic and unvaried they might be (see Robayo Acuña, 2025) – to meet real communication needs (Jones, 2022), learners still face challenges such as speaking anxiety, vocabulary gaps, and limited in-class opportunities (Fan & Yan, 2020; Mazdayasna, 2012). To illustrate, based on an assessment of students enrolled in an oral presentation course, Mazdayasna (2012) notes that students lacked structured feedback and time to organize ideas orally as well as clear criteria for oral assessment. In addition, language teachers frequently contend with large class sizes, limited instructional time, and an overemphasis on written skills, all of which restrict opportunities for scaffolded oral practice and personalized feedback (Galante et al., 2023).

Fernández-García and Fonseca-Mora (2022) examined the effect of emotional understanding and music as factors that might contribute to students' willingness to communicate and found that one way to boost students' motivation is to provide them with "meaningful communicative tasks about everyday topics" (p. 135). The persistent neglect of real-world tasks – in other words, authenticity (Jones, 2022) – they found in their study may be attributed to the usual focus on grammatical accuracy, an issue that should be reexamined as speaking does not rely on that dimension alone. Rather, the teaching of speaking should enable learners to use language meaningfully and confidently in social contexts (Fernández-García & Fonseca-Mora, 2022).

Sociocultural theories provide a lens through which speaking proficiency in a second or additional language can be understood not merely as the accumulation of discrete skills but as a socially-mediated process of meaning-making and identity construction. From this perspective, language learning is fundamentally interactional and dialogic and is shaped by the learner's engagement not only with the teacher but also with capable peers, tools, and contextual support. In line with this view, Schmid's (2017) work on teacher education in computer-assisted language learning (CALL) emphasizes the role of technology as a mediating factor, not as a replacement for instruction. Her CALL program highlights how, when embedded in reflective and collaborative classroom practices, digital tools can create conditions for scaffolded interaction and cognitive engagement. In the same vein, Neri et al. (2002, p. 441) argue that pedagogical constructs are often marginalized in favor of what the authors call "technical novelties" in their analysis of computer-assisted pronunciation training courseware. Neri et al. noted that technological limitations of the pronunciation training platforms they investigated are due to the very fact that their activities are not pedagogically informed, resulting in meaningless repetition exercises. Their study lists a number of suggestions, such as "includ[ing] pronunciation training within a comprehensive program based on a communicative approach" (Neri et al., 2002, p. 460); ensuring that models are more realistic and target at intelligibility rather than idealized norms; introducing tasks that allow for authentic practice instead of mechanical drills; adding more role-play activities that help students develop interactive skills; among others.

Nevertheless, studies have shown that teachers may also lack the knowledge to promote pedagogical activities that will help students enhance their oral skills. One such study is that of Robayo Acuña (2025), which investigated university English language teachers' understandings and practices of fluency in the English classroom. One of the study's participants, for example, mentioned that an activity he promotes in class is to have students use their own phones to record 2-minute speeches on a given topic to practice fluency. Placed under 'planning, rehearsal, and repetition', this kind of activity accounts for 70% of the fluency-oriented activities promoted by Robayo Acuña's participants. Robayo Acuña highlights the importance of research to inform pedagogical practices, as research can assist teachers in becoming better equipped for their students' needs. Thanks to the recent ample and easy access to technology such as AI chatbots, teachers may now afford to incorporate more meaningful oral practice in their classes, even though there has been some questioning that AI tools will replace those same teachers (Tavares et al., 2023). However, studies have drawn attention to the integration of AI tools in pedagogical practices rather than on the inevitable substitution of teachers by software.

Integrating AI chatbots in language education

The introduction of artificial intelligence (AI) into language education has brought about changes in how language learning may be optimized, individualized, and supported (Chiu, Xia, et al., 2023; Du & Daniel, 2024; Rudolph et al., 2023, 2025). Among these developments are AI-powered

chatbots, tools that generate texts through conversational robots (Artiles Rodríguez et al., 2021; Rudolph et al., 2023) and are increasingly being used to support language proficiency-oriented practice, particularly in academic writing. These tools provide a channel for real-time dialogue, offering nonjudgmental rehearsal spaces and providing feedback (Barrot, 2023a; Chen et al., 2025; Chiu, Moorhouse, et al., 2023; Chiu, Xia, et al., 2023; Du & Daniel, 2024).

Studies have shown that AI chatbots support autonomy by offering learners a safe and private environment in which to make mistakes and self-correct (Du & Daniel, 2024; Tai, 2024), thus increasing motivation and engagement when students are working with AI tools as these systems allow for repetition followed by customized feedback without social pressure (Tai, 2024). This aligns with sociocultural views of learning, in which technological tools serve as mediating elements that support learner development through guided interaction (Schmid, 2017).

Since teachers may devote a significant amount of class time to lecturing, they often find it challenging to find the time for drills or any other type of speaking practice in the classroom that would require individualized attention. Automating part of these tasks may be helpful in that AI tools provide them with technological assistance to both scaffold student interaction and monitor students' performance. To this end, teachers may employ AI chatbots to, for example, spot students who may require extra help so they can intervene early and offer guidance, thus helping to improve overall classroom performance (Baker & Smith, 2019).

Some examples of what those tools can do include metrics related to lexical range, syntactic complexity, or error frequency (Ranalli, 2018). Studies show that these tools allow for more personalized feedback than traditional teacher interventions but with limitations regarding concerns around algorithmic transparency, reliability, and learner trust (Zou et al., 2023). However, when framed as formative and supplemented by the teacher, these tools can foster reflection and awareness of individual learning processes (Baker & Smith, 2019; Ranalli, 2018). To illustrate, Zou et al. (2023) note that students tend to accept AI-generated evaluations when accompanied by explanatory feedback and framed as supplementary rather than definitive judgments.

In sum, the value of AI chatbots lies not only in their capacity to improve opportunities for language practice but also in their potential to foster critical AI literacy (Tseng & Warschauer, 2023; Warschauer et al., 2023). As Schmid (2017) emphasizes, language educators must integrate such tools along with critical awareness of the socio-institutional contexts in which they are embedded. Learners, in turn, should be encouraged to question the authority of automated feedback, understand how their data is used, and reflect on their interactions with these systems (Rudolph et al., 2023; Tseng & Warschauer, 2023; Warschauer et al., 2023). In this sense, incorporating chatbots in ELT classrooms becomes part of a broader agenda to prepare learners for critical, agentive participation in digital environments.

Integrating AI chatbots in speaking skill development

AI chatbots are increasingly being integrated into general language education, but their use in supporting speaking development does not seem to have followed the same pace as for writing. It could be explained by the fact that in ELT classrooms, speaking is often underemphasized due to large class sizes, exam-oriented curricula, and limited interaction time. Yet these very reasons justify the use of AI-based tools as these can assist teachers in providing learners with expanded opportunities for oral practice. However, designing interventions that truly enhance speaking skills rather than merely simulating conversation requires grounding in both learning theory and critical CALL perspectives.

As reviewed by Du and Daniel (2024), AI-powered tools that simulate spoken dialogue can help address the challenges learners face in ELT classrooms by providing them with contextualized speaking practice. However, their pedagogical value depends heavily on how they are integrated into the classroom since simply exposing learners to AI tools is insufficient (Li et al., 2025) and needs to be supplemented by guidance on the use of these tools for productive language output, reflection, and self-monitoring.

Zou et al. (2023) highlight the importance of learner perceptions and autonomy in the use of AI for speaking practice. Their study shows that students respond more positively to speech evaluation tools when the feedback is presented as formative and when learners retain agency over whether and how to act on it. Similarly, Ayedoun et al. (2019) found that learners benefit from chatbot-mediated speaking when tasks are clearly defined and connected to learners' communicative needs. These findings reinforce the idea that the integration of AI should involve the intentional sequencing of activities, framing the AI tool as a collaborator rather than an evaluator. Moreover, the process should involve tasks that help students reflect and transfer to real-world interaction (Li et al., 2025). Learners benefit most when tasks are contextualized, goal-oriented, and followed by feedback that helps them understand not only what went wrong but also how to improve their spoken performance. In a quasi-experimental study, Ahmad Al-husban (2025) observed that an AI-integrated pedagogical intervention helped the experimental group develop significantly in areas such as pronunciation, intonation, and vocabulary but not in grammar. That study also highlighted the importance of manipulation, that is, "providing students with detailed information about their current level of proficiency" (p. 123), demonstrating that learning awareness is key to students' progress.

In a meta-analysis of 15 studies addressing pedagogical work on pronunciation through AI-integrated instruction, Vančová (2023) found not only that such instruction assists in improving intelligibility but also that it helps students build confidence by lowering speaking anxiety. However, Vančová called for more studies of the effect of AI-integrated instruction on pronunciation, arguing that this remains an underrepresented area. Designing AI-supported speaking activities thus involves more than technical implementation. Rather, it calls for carefully

structured tasks, learner training in interpreting and using automated feedback, and opportunities for reflection on language, learning, and tool use. These principles inform the intervention described in this study, which explores how students interacted with AI chatbots during speaking practice in a higher education academic English context.

This study explores how AI-supported interventions can be implemented to support speaking development in an academic English university context. Drawing from a sociocultural perspective, the study considers not only linguistic outcomes but also learner agency and tool mediation in shaping the learning experience. We are guided by the following research questions:

1. What impact does the integration of Copilot AI have on B1 university students' speaking proficiency (including fluency, accuracy, pronunciation, and vocabulary) compared with traditional instruction?
2. How do students perceive the use of Copilot AI for speaking practice, and in what ways does teacher mediation shape their engagement, confidence, and use of feedback during the intervention?

Methodology

To examine the extent to which the Copilot AI application (https://copilot.microsoft.com) affects the speaking proficiency of English language learners, this study adopted a quasi-experimental design consisting of a control group and an experimental group. What is explored here is the pedagogical intervention applied to a group of English language students at a private university in Georgia, which followed a pilot study within the same institution. The pilot study yielded aspects such as (1) the importance of scaffolding both by other students and teachers, (2) close monitoring of the AI chatbot's outcomes given that it may present inconsistency in controlling the proficiency level, (3) the inclusion of proficiency level in the prompt that students gave to the AI chatbot. These were taken into account in the design of the activities that were delivered to the experimental group, described below.

The context

Participants in the experiment were 40 undergraduate students aged 17 to 19 at a private university in Georgia, who all placed at B1 proficiency level (Common European Framework of Reference (CEFR)) through a test delivered at the beginning of the course. Participants were informed of all aspects of the study and gave written consent. The experiment lasted approximately one semester (i.e., 12 weeks), in two weekly classes consisting of 2 hours each at the start of the Spring 2025 semester. The sample was drawn from a General English class designed for freshmen from

different faculties. The students were divided into two groups: a control group with 20 students and an experimental group with 20 students (it originally started with 21 students, but one student unfortunately withdrew from the course). Of the four weekly contact hours, one hour was reserved for the pedagogical intervention with the Copilot AI chatbot reported here only with the experimental group. Lesson planning followed the university's assigned syllabus for English language teaching, with the inclusion of worksheets designed to implement the integration of the Copilot AI application as a scaffolding tool.

Pedagogical intervention

The classes followed the topics (i.e. health, traveling) suggested in the syllabus. The textbook adopted was *Roadmap B1* (Maris & Williams, 2020), published by Pearson Education. The chosen AI tool that assisted with the experimental group was Microsoft Copilot, since it was available for free at the time of the intervention. The tool also afforded the use of speech through a microphone, with an oral response. This could be done with the use of students' personal mobile devices. Additionally, the tool could keep the written record of the complete conversation records, which were later collected from the students and combined with the reflections for further analysis.

To support systematic integration of the AI chatbot with the experimental group, we designed a Collaborative Model (Figure 1).

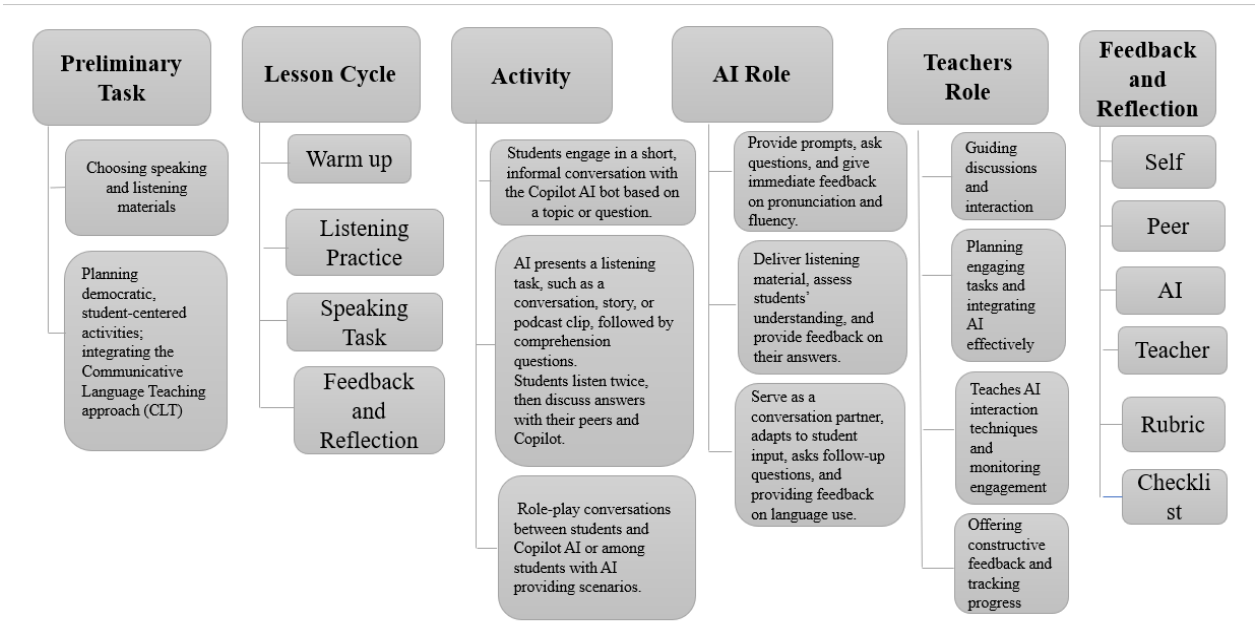

Figure 1. The Collaborative Model.

Each lesson was structured into three phases: (1) Planning, (2) Implementation, and (3) Feedback. The model was intentionally simple so that it could be replicated by teachers with varying levels of technological experience or institutional support. This model created a complementary learning environment in which the teacher had the role of guiding and monitoring the students' progress with the control group, or shared that role with the AI tool in terms of ensuring speaking practice and immediate feedback for each student in the experimental group.

In the planning phase, the teacher selected a communicative speaking task aligned with the course syllabus and adapted prompts to ensure the chatbot, particularly during the feedback phase, operated at an appropriate proficiency level (B1). The teacher's instructions, scaffolding questions, and expected interaction sequences were incorporated into worksheets. During the implementation phase, the experimental group used Copilot for controlled and semi-controlled speaking tasks (such as warm-up dialogues, role plays, or simulated interviews). The teacher monitored interaction, clarified misunderstandings, and intervened when the AI provided inaccurate or inappropriate output (for example, if the output was too advanced for the students' level). The control group completed parallel tasks without technology, using teacher-led modelling and peer interaction. In the feedback phase, students were invited to engage in reflection (oral or written), completed a self-assessment checklist, and compared teacher, peer, and AI feedback. Every three to four weeks, the teacher proposed a reflection on the AI-integrated activities the students had engaged in, with the teacher taking notes while the students presented their ideas. The reflections, along with the interviews conducted with the students at the end of the course, were collected and qualitatively analyzed.

Instruments

The experiment started with a pre-test for both groups, consisting of speaking tasks designed to identify the students' proficiency level. Their progress along with their reflections during the course was monitored by integrating a continuous observation sheet. A post-test, which mirrored the tasks of the pre-test, was conducted at the end of the semester.

To strengthen the reliability of the speaking assessment, the pre- and post-test performances were independently evaluated by three raters with experience in ELT. Prior to scoring, the raters participated in a short calibration session during which the evaluation rubric (Appendix A) was reviewed and applied to a small set of sample recordings in order to ensure shared interpretation of the scoring criteria. Each rater scored the performances individually using the rubric provided in the course syllabus, which assessed fluency, pronunciation, grammar, and vocabulary (Fulcher, 2014) through units (Luoma, 2004). After the initial scoring stage, the ratings were compared to examine the level of agreement among evaluators. Inter-rater reliability was calculated using correlation analysis across raters' scores. The resulting agreement levels were high, indicating

consistent application of the rubric and strengthening the reliability of the speaking proficiency measurements used in the study. In cases where discrepancies between raters exceeded one band on the rubric scale, the scores were reviewed and discussed until a consensus score was reached.

Another instrument used was the self-assessment checklist (Table 1), given to each student in the experimental group, and which served as a basis for them to monitor their progress. It was also used in the reflection on AI-integrated activities. The latter was enriched by an analysis of the recorded scripts collected from all students in each activity.

Table 1. Self-assessment checklist.

Self-assessment checklist	
( )	I participated in a warm-up activity with Copilot AI, teachers, and peers.
( )	I listened to the material twice.
( )	I completed the listening task.
( )	I compared my answers with Copilot AI and peers.
( )	I engaged in role-play conversations with Copilot AI and peers.
( )	I received immediate feedback from Copilot AI, peers, and a teacher.
( )	I noted feedback in my notebook (paper-based, digital).
( )	I reflected on my performance and assessed my progress in listening and speaking.
( )	I was given the assessment rubrics to assess my progress.
( )	I received delayed feedback from my teacher.
( )	I completed all the items on this checklist.

Data analysis

Statistical analyses were performed using SPSS 16.0 for both the pre- and post-tests for each group. We first calculated the mean, mode, median, standard deviation, skewness, and kurtosis for all students' responses (n = 40) from the pre- and post-tests. We then compared each group's performance. Finally, we conducted independent t-tests to determine whether any differences between the experimental and control groups were statistically significant.

The teacher's notes taken during in-class discussions were analyzed qualitatively and compared with the analysis of both pre- and post-tests.

Ethical considerations

The intervention followed ethical principles set out at two different universities: the university where the data were collected and the university where the study was conducted. After the participants were informed about the research procedure, they gave written consent. All data were anonymized during the data analysis.

Analysis

Referring back to the Research Question 1 (What impact does the integration of Copilot AI have on B1 university students' speaking proficiency, including fluency, accuracy, pronunciation, and vocabulary, compared with traditional instruction?), we now turn to a discussion of the results of the pre- and post-tests of both groups to examine whether the intervention influenced the students' speaking practice. Next, to answer Research Question 2 (How do students perceive the use of Copilot AI for speaking practice, and in what ways does teacher mediation shape their engagement, confidence, and use of feedback during the intervention?), we present a summary of the results of the post-test interpreted through a qualitative analysis of the students' perceptions through their self-assessments and the teacher's notes.

Table 2 shows the results of the pre- and post- speaking tests and their respective progress indicators. Different codes were created for each individual participant to ensure that these reflected the systematic nature of the data.

Table 2. Experimental group: Pre- and post-test results.

Students	Pre-Test	Post-Test	Improvement
EXP_S_1	3	5	+2
EXP_S_2	5	8	+3
EXP_S_3	4	6	+2
EXP_S_4	7	10	+3
EXP_S_5	8	10	+2
EXP_S_6	9	10	+1
EXP_S_7	2	5	+3
EXP_S_8	5	9	+4
EXP_S_9	4	6	+2
EXP_S_10	3	6	+3
EXP_S_11	4	7	+3
EXP_S_12	2	5	+3
EXP_S_13	3	5	+2
EXP_S_14	4	9	+5
EXP_S_15	4	8	+4
EXP_S_16	5	7	+2

EXP_S_17	2	4	+2
EXP_S_18	1	3	+2
EXP_S_19	1	3	+2
EXP_S_20	4	6	+2
Mean	**4.00**	**6.60**	
Median	**4.00**	**6.00**	
Standard Deviation	**2.05**	**2.28**	

The results show consistent positive improvements among students over the course of the experiment. The most common improvement (+2 points) was achieved by 45% (N=9) of students, while students with middle-range pre-test scores demonstrated the largest gains (+4 and +5 points). Results from the speaking pre-test revealed a wide range of initial proficiency levels, which may have influenced the degree of improvement observed in the post-test.

Table 3 shows the correlation analysis of students' pre- and post-speaking scores. The analysis included data from the 20 students who worked with the AI tool. The Pearson correlation coefficient was 0.862, a high correlation level (over 0.8) widely accepted as showing a strong positive relationship between two variables (here, VAR00001 (pre-test) and VAR00002 (post-test)). The significance level was p =< 0.001, suggesting that the correlation between the two groups is statistically significant. In other words, students who scored higher on the pre-test also tended to score higher on the post-test, with the results unlikely to be due to chance.

Table 3. Correlation between pre-test and post-test speaking scores in the experimental group.

		VAR00001	VAR00002
VAR00001	Pearson Correlation	1	.862
	Sig. (2-tailed)		< .001
	N	20	20
VAR00002	Pearson Correlation	.862*	1
	Sig. (2-tailed)	< .001	
	N	20	20

* Correlation significant at the 0.01 level (2-tailed).

Statistical analysis evidences a feature that we observed in the pre-test results, showing that most of the students found it difficult to speak freely on the topics under discussion, and frequent

grammatical errors sometimes made sentences difficult and incomprehensible. On the other hand, the post-test results demonstrate that students felt more confident and spoke at length.

In the pre-test, pronunciation was the criterion with the lowest score among all participants. In relation to fluency, some students made long unfilled pauses while speaking, while some others spoke at length but struggled with their vocabulary use – even though vocabulary use per se did not prevent understanding. All areas of pronunciation, fluency and vocabulary indicated an improvement in the results from the post-test. This improvement may result from each student's engagement and motivation observed throughout the whole experiment. Ongoing observation and analysis of the students' self-assessment checklists showed that, in the experimental group, there was a significant increase in attendance and students' participation. This increased participation, we believe, was likely due to the excitement behind the pedagogical intervention which contributed to the improvement in their speaking as engaged participants.

Table 4. Control group: Pre- and post-test results.

Students	Pre-Test	Post-test	Improvement
CON_S_1	5	6	+1
CON_S_ 2	1	3	+2
CON_S_3	6	7	+1
CON_S_4	1	4	+3
CON_S_ 5	4	5	+1
CON_S_ 6	4	7	+3
CON_S_ 7	5	7	+2
CON_S_ 8	7	10	+3
CON_S_ 9	1	3	+2
CON_S_ 10	4	6	+2
CON_S_ 11	4	5	+1
CON_S_ 12	2	5	+3
CON_S_ 13	7	7	+0
CON_S_ 14	7	8	+1
CON_S_ 15	7	8	+1
CON_S_ 16	3	4	+1
CON_S_ 17	1	3	+2
CON_S_ 18	4	4	+0

CON_S_ 19	4	5	+1
CON_S_ 20	5	9	+2
Mean	**4.10**	**5.80**	
Median	**4.00**	**5.50**	
Standard Deviation	**2.10**	**2.04**	

Over the course of the study, the control group displayed moderate increases in speaking scores, most students showing small incremental improvements in each assessment. The typical improvement for students in the control group was +1 point. Those control group students who improved by +2 or +3 points represented approximately 35% (N=7) of the group, with two students recording no improvement. Overall, the average score rose from 4.10 (pre-test) to 5.80 (post-test), and the median score from 4.00 to 5.50. Although the control group demonstrated improvements in scores, compared with the experimental group, improvement was less pronounced.

Table 5. Correlation between pre-test and post-test speaking scores in the control group.

Correlations			
		VAR00001	VAR00002
VAR00001	Pearson Correlation	1	.95**
	Sig. (2-tailed)		<.001
	N	20	20
VAR00002	Pearson Correlation	.95**	1
	Sig. (2-tailed)	<.0001	
	N	20	20

** Correlation significant at the 0.01 level (2-tailed).

Table 5 shows the correlation analysis of the pre-test and post-test scores of 20 students in the control group. The Pearson correlation coefficient was 0.95, indicating a strong positive relationship between the two variables: VAR00001 (pre-test) and VAR00002 (post-test). The significance level was $p < .001$, indicating a correlation that was not only high but also statistically significant.

Table 6. Comparison between the experimental and control groups.

Metric	**Experimental Group (With AI)**	**Control Group (Without AI)**
Pre-Test Mean	4	4.1

Post-Test Mean	6.6	5.8
Mean Improvement	+2.60	+1.70
Pre-Test Median	4	4
Post-Test Median	6	5.5
Standard Deviation (Pre-)	2.05	2.1
Standard Deviation (Post-)	2.28	2.04
Most Common Improvement	+2 (45% of students)	+1 (typical)
Max. Individual Gain	+5	+3
Min. Individual Gain	+1	0
Pearson Correlation	0.862	0.95
Significance (*p*-value)	< .001	< .001
Interpretation of Correlation	Strong Positive	Very Strong Positive

Table 6, with a comparison between the experimental and control groups, shows notable differences in learning outcomes. Both groups demonstrated improvement in speaking scores from pre-test to post-test. However, the experimental group, which engaged with the AI chatbot, exhibited a higher mean improvement (+2.60) compared to the control group (+1.70). The experimental group also recorded the highest individual gain (+5), whereas the maximum for the control group was +3. While the control group included two participants who showed no improvement, all students in the experimental group demonstrated progress. Pearson correlation results indicated strong positive relationships between pre- and post-test scores for both groups (experimental: 0.862; control: 0.95), reflecting consistency in performance rankings across tests. Nonetheless, the experimental group achieved higher post-test means and demonstrated greater gains in fluency, pronunciation, and vocabulary.

Based on the post-test results, the students' self-reflections and classroom observations of the experimental group, we have identified four main areas in which students' achievements were more significant. The areas, Grammar, Vocabulary, Pronunciation and Confidence, are listed in Table 7, along with a brief summary of achievements.

Table 7. Students' achievements.

Progress	Indicators
Grammar	Students began using articles, prepositions, and basic grammatical tenses such as the present simple, present continuous, past simple, and past continuous more effectively for their communicative goals. The frequency of previously common errors such as the omission of auxiliary verbs in negative and interrogative forms was substantially reduced.
Vocabulary	Students incorporated a wider range of words and phrases into their speech, showcasing a more diverse and resourceful vocabulary.
Pronunciation	Many students, who initially pronounced words as they appeared in writing (e.g., "every" pronounced exactly as it is spelled, i.e., as three syllables), showed marked improvements in stress and vowel reduction. Their pronunciation became more consistently intelligible.
Confidence	Many students appeared to partially overcome hesitation in speaking, and some reported increased confidence. Not all students showed the same degree of change, but several no longer hesitated to express themselves in certain contexts, suggesting progress in reducing speech-related anxieties.

These improvements, as mentioned above, can be attributed to students' high levels of motivation and engagement during the experiment, but also to the extra support that each student in the experimental group received through the AI chatbot. We turn to a more thorough discussion below.

Discussion

The pre-designed activities delivered by the teacher with the integration of AI tools granted the students the opportunity to enhance their language skills, particularly in grammar, vocabulary, and pronunciation. This improvement partially corroborates the study of Ahmad Al-husban (2025), which noted improvements in vocabulary and pronunciation, but did not observe the same evolution in grammar. In contrast, our participants showed an enhanced understanding of key grammatical structures such as the correct use of articles, prepositions, and auxiliary verbs in negative and interrogative forms. While the study did not isolate grammar as an independent experimental variable, the convergence of rubric scores and qualitative observations suggests that repeated interaction with the AI chatbot indeed provided opportunities for noticing and self-correction that may have contributed to improvements in grammatical accuracy. This then suggests that the personalized feedback and modeling provided by the AI chatbot allowed the learners to self-correct and internalize rules more effectively than in traditional classroom settings, as seen in Du and Daniel (2024) and Tai (2024). Regarding vocabulary, exposure to a wider range of language input that extrapolated those from the textbook may have facilitated the acquisition (and retrieval) of new words and phrases, aligning with research that highlights the importance of diverse and rich linguistic environments in language acquisition (Fernández-García & Fonseca-

Mora, 2022; Tai, 2024; Neri et al., 2002; Robayo Acuña, 2025). The contextualized interactions allowed learners to encounter target vocabulary more often across different situations, reinforcing retention through repetition and relevance. As with pronunciation, the immediate feedback offered by the chatbot helped students correct their mistakes as they repeatedly heard and mimicked readable pronunciation in real-time, further supporting not only accuracy but also their self-confidence (Vančová, 2023). We noticed that, as students felt free of pressure from (human) judgment, this practice likely contributed to a more relaxed and productive learning environment, helping students refine their phonological skills.

Besides the improvement in accuracy and intelligibility, the interactions with the AI chatbot seemed to have enriched the students' learning experience. As observed in the teacher's notes, the AI tool assisted in bridging the gap between the textbook and real conversational use, allowing for extra practice that could not be allocated in the control group. Through practice with AI-simulated conversation partners, students were encouraged to refine their speaking skills in a low-pressure environment, thus improving their real-world interaction abilities. This in turn may have helped students overcome psychological barriers to speaking, such as fear of making mistakes or feeling judged (Fernández-García & Fonseca-Mora, 2022). The reduction in speech-related anxiety and the boost in self-confidence seen in the fluency improvement highlights the role of the AI chatbot as a supportive, non-judgmental conversational partner. This has been shown to be helpful in language learning as confidence plays a significant role in language production and the willingness to practice (Fernández-García & Fonseca-Mora, 2022).

In the recurrent reflection sessions, the students noted that the teacher was allowed to allocate more time to individualizing the teaching (Du & Daniel, 2024), having more time to monitor and scaffold the activities. But most importantly, the students highlighted potential applications of the AI tool outside the classroom, demonstrating a growing sense of agency. Some students were motivated to continue practicing spoken English outside school hours, thus enacting self-directed language learning. Others reported using the chatbot during their free time, having casual conversations with it, or returning to structured exercises, which were helpful for extra reinforcement and practice. This prolonged exposure to the target language provided an opportunity to continue practicing their new skills in a more relaxed setting. It also led the students to see communicating in English not just as a classroom obligation but rather as a useful and fun skill they could use in various settings. Unlike motivation, which reflected their desire to learn, this growing sense of agency involved making choices: deciding when to practice, which topics to explore with the chatbot, and how to evaluate their own progress. Not all students, however, reached this level of self-direction, but for those who did, the AI tool served less as a tutor and more as a resource they learned to control. Additionally, the students expressed their eagerness to follow the proposed Collaborative Model because, according to them, it brought the real world into the classroom. Through practice with the AI chatbot, they felt more motivated to speak in the target language. They claimed that the classroom environment developed into a secure place where

the combination of the teacher and the AI tool collaboratively contributed to their academic achievement, motivation, and self-confidence.

These results suggest that incorporating AI as a collaborative element in language learning not only improved students' speaking proficiency but also helped create a more supportive learning community as students could work with the teacher, their peers, and the chatbot, or with all three players at various times. This dynamic interaction between teacher and students mediated by the AI tool enhanced the learners' self-confidence and autonomy, as seen in Schmid (2017). Students became more proactive about participating in discussions or starting conversations in class, during which they responded with less hesitation and greater spontaneity. With the progression of the intervention, the teacher began to move away from a teacher-centered perspective and towards a learner-centered approach, with technology facilitating personalized learning and real-time feedback, as also suggested in Schmid (2017). This combined teamwork created a positive atmosphere in which meaningful speaking practice could take place.

However, these findings should be taken with caution. It is not our objective to promote blind, uncritical techno-enthusiasm or a technologization of education (Guilherme, 2017; Rudolph et al., 2023, 2025). The pedagogical intervention proposed here departed from studies of the development of English-speaking proficiency from a sociocultural perspective, in which human relations – whether student-student or teacher-student – are prioritized, as seen in our proposed Collaborative Model. The framework suggested here emphasizes the AI chatbot as a tool at a certain stage in the lesson, not as a substitute for the teacher. Therefore, teachers should be critical of the use of AI tools, just as with the use of any technological device. In classes in which each student has limited opportunities to speak, chatbots may indeed offer an additional interlocutor that helps optimize the class time, allowing for extended practice. This, however, may not reduce teacher workload, as planning requires more attention: teachers should provide task-specific prompts, define the role of the chatbot (interviewer, travel agent, friend), and specify the target level. Additionally, even though students reported reduced speaking anxiety after practicing first with the chatbot, this work needs to be monitored by teachers; the chatbot seemed to focus more on general language patterns such as lexical range, pronunciation, and grammatical structures, but the teacher had to provide interpretation and guidance about how to act on AI-generated suggestions.

We should note that another advantage of this intervention is its low-resource design. The AI chatbot used in this study (Microsoft Copilot) was freely accessible and our students could operate it entirely on their mobile devices, with no additional software, institutional licensing, or laboratory facilities required. The Collaborative Model relies on simple worksheets and structured prompts, making it thus feasible in contexts with limited technological infrastructure or restricted budgets. This is particularly relevant for settings in which teachers face large classes, limited time, and minimal or no access to language labs or paid platforms. We insist: while AI chatbots can serve as

valuable tools for scaffolding speaking practice, their integration demands a reassertion – not a reduction – of the role of the teacher. In the intervention described here, the chatbot acted as a supplement to human interaction, not a replacement. This distinction is important as the teacher remains the architect of the learning environment, designing activities, mediating the limitations of AI, and fostering critical engagement.

Several limitations should be acknowledged. First, the study was conducted with a relatively small sample from a single private university in Georgia, which may restrict the generalizability of the findings. Second, the intervention lasted only twelve weeks; longer-term studies would be needed to examine retention and sustained effects on speaking development. Third, although the teacher carefully monitored the output of the chatbot, it occasionally produced higher level language or inconsistent feedback, which could have influenced learner performance. As the syllabus dictated the sequence of topics, speaking tasks were limited to textbook themes rather than a wider range of authentic communicative contexts. Furthermore, the level of subjects' curiosity about using a new and unfamiliar tool may have slightly affected their engagement and academic performance, rather than the pedagogical framework itself. The motivation bias may have had an impact, as the subjects using this innovative AI chatbot could represent the group with already established motivation toward language learning. The longer period of using the AI chatbot technology for speaking purposes with larger, more diverse samples will make its use a routine and enable purely measuring the effects of the pedagogical framework. Future research could address these limitations by adopting longitudinal designs, larger samples, and multiple institutional contexts.

Conclusion and recommendations

Designed to incorporate an AI chatbot in English-speaking classes, the Collaborative Model, our findings suggest, may support the development of English language speaking skills among university students. Integrating Copilot AI as a teaching assistant in the real classroom helped shape a student-centered approach to teaching, leading to improvement in academic performance, motivation, and self-confidence. This analysis shows the potential of a collaborative learning modality in which the AI tool served as a pedagogical assistant to a human teacher. Besides the statistical data revealing measurable outcomes, the implications of our findings are broader, including the conclusion that integrating AI tools into the classroom has the potential to change conventional dynamics towards greater autonomy, communicative competence, and decreased affective barriers to language practice and thus, language use.

Importantly, these results contribute to and build on the established literature that encourages the inclusion of intelligent tutoring systems and digital instruments in language education (Ahmad Al-husban, 2025; Baker & Smith, 2019; Chiu, Moorhouse et al., 2023; Chiu, Xia et al., 2023; Li et al., 2025; Tai, 2024). Though previous research has established that AI can provide immediate

feedback and expose English language learners to a variety of language input, our study adds to the literature by showing that a sustained classroom-based implementation of AI needs to be done with teacher engagement. Far from suggesting that AI could replace educators, these results underscore the role of teachers in contextualizing learning culturally and emotionally and in delivering the individualized guidance AI systems lack. When purposefully incorporated in instruction, the AI tool can become a supplement to human teaching, enhancing — not replacing — the process of human learning.

References

Ahmad Al-husban, N. (2025). The impact of AI-assisted language learning tools on augmenting university EFL students' speaking skills in Jordan. *Journal of Applied Learning & Teaching*, *8*(1). https://doi.org/10.37074/jalt.2025.8.1.13

Artiles Rodríguez, J., Guerra Santana, M., Aguiar Perera, V., & Rodríguez Pulido, J. (2021). Agente conversacional virtual: La inteligencia artificial para el aprendizaje autónomo. *Pixel-Bit, Revista de Medios y Educación,* (62), 107–144. https://doi.org/10.12795/pixelbit.86171

Ayedoun, E., Hayashi, Y., & Seta, K. (2019). Adding communicative and affective strategies to an embodied conversational agent to enhance second language learners' willingness to communicate. *International Journal of Artificial Intelligence in Education*, *29*(1), 29–57. https://doi.org/10.1007/s40593-018-0171-6

Baker, T., & Smith, L. (2019). *Educ-AI-tion rebooted? Exploring the future of artificial intelligence in schools and colleges*. Nesta. https://media.nesta.org.uk/documents/Future_of_AI_and_education_v5_WEB.pdf

Barrot, J. S. (2023a). Using automated written corrective feedback in the writing classrooms: Effects on L2 writing accuracy. *Computer Assisted Language Learning*, *36*(4), 584–607. https://doi.org/10.1080/09588221.2021.1936071

Barrot, J. S. (2023b). Using ChatGPT for second language writing: Pitfalls and potentials. *Assessing Writing*, *57*, 100745. https://doi.org/10.1016/j.asw.2023.100745

Chen, X., Zhou, Z., & Prado, M. (2025). ChatGPT-3.5 as an automatic scoring system and feedback provider in IELTS exams. *International Journal of Assessment Tools in Education*, *12*(1), 62–77. https://doi.org/10.21449/ijate.1496193

Chiu, T. K. F., Moorhouse, B. L., Chai, C. S., & Ismailov, M. (2023). Teacher support and student motivation to learn with Artificial Intelligence (AI) based chatbot. *Interactive Learning Environments*, 1–17. https://doi.org/10.1080/10494820.2023.2172044

Chiu, T. K. F., Xia, Q., Zhou, X., Chai, C. S., & Cheng, M. (2023). Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence

in education. *Computers and Education: Artificial Intelligence*, *4*, 100118. https://doi.org/10.1016/j.caeai.2022.100118

Du, J., & Daniel, B. K. (2024). Transforming language education: A systematic review of AI-powered chatbots for English as a foreign language speaking practice. *Computers and Education: Artificial Intelligence, 6*, 100230. https://doi.org/10.1016/j.caeai.2024.100230

Fan, J., & Yan, X. (2020). Assessing speaking proficiency: A narrative review of speaking assessment research within the argument-based validation framework. *Frontiers in Psychology*, *11*, 330. https://doi.org/10.3389/fpsyg.2020.00330

Fernández-García, A., & Fonseca-Mora, M. C. (2022). EFL learners' speaking proficiency and its connection to emotional understanding, willingness to communicate and musical experience. *Language Teaching Research*, *26*(1), 124–140. https://doi.org/10.1177/1362168819891868

Fulcher, G. (2014). *Testing second language speaking*. Routledge. https://doi.org/10.4324/9781315837376

Galante, A., Zeaiter, L. F., Dela Cruz, J. W. N., Massoud, N., Lee, L., Aronson, J., De Oliveira, D. S. A., & Teodoro-Torres, J. A. (2023). Digital plurilingual pedagogies in foreign language classes: Empowering language learners to speak in the target language. *Language Learning Journal*, *51*(4), 523–543. https://doi.org/10.1080/09571736.2023.2179654

Guilherme, A. (2017). AI and education: The importance of teacher and student relations. *AI & Society*, *34*(1), 47–54. https://doi.org/10.1007/s00146-017-0693-8

Jones, C. (2022). Authenticity in language teaching materials. In J. E. Norton & H. Buchanan (Eds.), *The Routledge handbook of materials development for language teaching* (pp. 65–77). Routledge.

Li, Y., Wu, Y., & Chiu, T. K. F. (2025). How teacher presence affects student engagement with a generative artificial intelligence chatbot in learning designed with first principles of instruction. *Journal of Research on Technology in Education*, 1–17. https://doi.org/10.1080/15391523.2025.2493942

Luoma, S. (2004). *Assessing speaking*. Cambridge University Press.

Mazdayasna, G. (2012). Objective assessment of oral presentations and EFL learners' speaking development. *Journal of Foreign Language Teaching and Translation Studies*, *1*(1), 23–38. https://doi.org/10.22034/EFL.2012.79159

Maris, A., & Williams, D. (2020). *Roadmap B1: Student's Book*. Pearson Education Limited.

Neri, A., Cucchiarini, C., Strik, H., & Boves, L. (2002). The Pedagogy-Technology Interface in Computer Assisted Pronunciation Training. *Computer Assisted Language Learning*, *15*(5), 441–467. https://doi.org/10.1076/call.15.5.441.13473

Ranalli, J. (2018). Automated written corrective feedback: How well can students make use of it? *Computer Assisted Language Learning*, *31*(7), 653–674. https://doi.org/10.1080/09588221.2018.1428994

Robayo Acuña, L. M. (2025). Developing Oral Fluency in English: EFL Teachers' Understanding, Self-Reported Practices, and Textbook Support. *Profile: Issues in Teachers' Professional Development, 27*(2), 121–136. https://doi.org/10.15446/profile.v27n2.115146

Rudolph, J., Ismail, F., Tan, S., & Seah, P. (2025). Don't believe the hype. AI myths and the need for a critical approach in higher education. *Journal of Applied Learning & Teaching*, *8*(1), 6–27. https://doi.org/10.37074/jalt.2025.8.1.1

Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning & Teaching*, *6*(1). https://doi.org/10.37074/jalt.2023.6.1.9

Schmid, E. C. (2017). *Teacher education in computer-assisted language learning: A sociocultural and linguistic perspective*. Bloomsbury Academic.

Tai, T.-Y. (2024). Effects of intelligent personal assistants on EFL learners' oral proficiency outside the classroom. *Computer Assisted Language Learning*, *37*(5–6), 1281–1310. https://doi.org/10.1080/09588221.2022.2075013

Tavares, C., Oliveira, L., Duarte, P., & Da Silva, M. M. (2023). Artificial intelligence: A blessing or a threat for language service providers in Portugal. *Informatics*, *10*(4), 81. https://doi.org/10.3390/informatics10040081

Tseng, W., & Warschauer, M. (2023). AI-writing tools in education: If you can't beat them, join them. *Journal of China Computer-Assisted Language Learning*, *3*(2), 258–262. https://doi.org/10.1515/jccall-2023-0008

Vančová, H. (2023). AI and AI-powered tools for pronunciation training. *Journal of Language and Cultural Education, 11*(3), 12–24. https://doi.org/10.2478/jolace-2023-0022

Warschauer, M., Tseng, W., Yim, S., Webster, T., Jacob, S., Du, Q., & Tate, T. (2023). The affordances and contradictions of AI-generated text for writers of English as a second or foreign language. *Journal of Second Language Writing*, *62*, 101071. https://doi.org/10.1016/j.jslw.2023.101071

Zou, B., Lyu, Q., Han, Y., Li, Z., & Zhang, W. (2023). Exploring students' acceptance of an artificial intelligence speech evaluation program for EFL speaking practice: An application of the integrated model of technology acceptance. *Computer Assisted Language Learning*, 1–26. https://doi.org/10.1080/09588221.2023.2278608

Appendix A: Evaluation rubric

Points	Fluency	Pronunciation	Grammar	Vocabulary	Details
9-10	Smooth and fluid speech; few to no hesitations; volume is excellent.	Pronunciation is excellent.	Excellent control of language features; a wide range of well-chosen vocabulary.	Accuracy and variety of grammatical structures.	High level of accuracy in description is reflected; high degree of detail included in description.
7-8	Smooth and fluid speech; few hesitations; slight search for words; inaudible word or two.	Pronunciation is good.	Good language control; good range of relatively well-chosen vocabulary.	Some errors of grammatical structures possibly caused by attempt to include variety.	High level of accuracy in description is reflected; more detail should be included.
5-6	Speech is relatively smooth but is characterized by some hesitation and unevenness caused by rephrasing and searching for words; volume wavers.	Pronunciation is adequate. No miscommunication due to mispronunciation.	Adequate language control; vocabulary is lacking.	Frequent grammatical errors that do not obscure meaning; little variety of structures.	Good accuracy in description, though some detail might be lacking.
3-4	Speech is frequently hesitant and jerky, with some sentences incomplete; volume very soft.	Pronunciation is adequate. Miscommunication may occur, but not of essential information.	Weak language control; basic vocabulary choice, with some words clearly lacking.	Grammatical errors even in simple structures that obscure meaning.	Description lacks some accuracy and some critical details are missing that makes it difficult to understand.
1-2	Speech is slow and exceedingly hesitant and strained except for short, memorized phrases.	Pronunciation is lacking and hard to understand.	Weak language control; vocabulary that is used does not match the task.	Frequent grammatical errors even in simple structures that obscure meaning.	Description is lacking.

Eliza Kintsurashvili is a certified English language teacher and experienced educator from Georgia with a strong academic background, including multiple degrees and a recently completed PhD. She holds the internationally recognized CELTA certificate and has undertaken extensive professional training, including TEFL/TESOL, Leaders of Learning (Harvard University), Training of Trainers, and courses in e-learning and English language teaching, reflecting her strong commitment to continuous professional development and excellence in education.

Ekaterine Pipia is a Professor at the International Black Sea University, known for her expertise in educational sciences and her contributions to the fields of teaching methodologies and educational research. She has published extensively on topics such as curriculum development, innovative teaching practices, and intercultural communication in education. Her dedication to mentoring students and supporting emerging researchers has contributed significantly to the enrichment of educational practices in Georgia.

Malila Prado is an Assistant Professor in the Department of Languages and Cultures at Beijing Normal-Hong Kong Baptist University (BNBU), China. She has worked as an English language teacher for 25 years. She holds a PhD from the University of Sao Paulo (Brazil), focusing on corpus-based analyses of aviation communication. Her current research examines pragmatic competence, corpus-based material design, and the pedagogical use of AI to support multilingual learners' production.