IDEAS

Journal on Language Teaching and Learning, Linguistics and Literature

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

Volume 14, Number 2, December 2026
pp. 9149 - 9164

Copyright © 2026 The Author
IDEAS is licensed under CC-BY-SA 4.0 License



Issued by English study program of UIN Palopo

The Effect of AI-Assisted Preparation on Students' Speaking Performance in Monologue Report Text Tasks

Fadliyah Uzma[1], Meidy Kautsar Arriv[2], Gelar Dwirahayu[3], Ismalianing Eviyuliwati[4], Alek[5]

[1,2,3,4]Universitas Islam Negeri Syarif Hidayatullah Jakarta, Indonesia

Article Info

Received: 2026-06-25
Revised: 2026-10-28
Accepted: 2026-10-29

Keywords:

AI-assisted preparation;
EFL speaking;
monologue;
report text;
ChatGPT;

DOI:

10.24256/ideas. v14i2.11455

Corresponding Author:

Fadliyah Uzma

fadliyahuzma96@gmail.com

Universitas Islam Negeri Syarif Hidayatullah Jakarta

Indonesia

Abstract

Although artificial intelligence (AI) is increasingly used in EFL speaking instruction, little is known about its role as a pre-speaking preparation tool for genre-based monologue tasks. This study investigates the effect of AI-assisted speaking preparation on the monologue speaking performance of 9th-grade students in the context of English as a Foreign Language (EFL). Employing a mixed methods design, this study integrated a pre-experimental one-group pre-test–post-test design with qualitative content analysis of student prompts used during AI-assisted preparation. A total of 20 students from a Madrasah Tsanawiyah in Bekasi participated as the single group. Quantitative data were collected through pre-test and post-test assessments of monologue speaking performance in an animal-themed report text task. Quantitative results indicated a statistically significant improvement in speaking scores ($t = -9.790$, $df = 19$, $p < .001$), with an overall mean N-Gain of 0.25, categorized as low. Qualitative analysis of the prompts of three students suggested that more specific and strategic prompts were associated with greater speaking improvement, although this pattern is illustrative and needs confirmation in larger samples. Because the study used a one-group design without a control group, the results indicate that speaking scores improved after AI-assisted preparation, not that AI alone caused the improvement. The findings suggest that how students interact with AI may matter as much as having access to it, and that explicit instruction in prompt formulation may be needed.

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

1. Introduction

The integration of artificial intelligence (AI) into education has expanded rapidly across disciplines, with language learning and teaching emerging as one of the most dynamic areas of application (Zawacki-Richter et al., 2019). AI tools now offer features that make the learning process more engaging, interactive, and adaptive to individual learner needs (Chen et al., 2020; Huang et al., 2023). In the domain of English as a Foreign Language (EFL), the use of AI has grown substantially as a language learning medium, particularly in skills that require real-time production such as speaking (Haristiani, 2019; Huang et al., 2023). Among the specific benefits attributed to AI-assisted learning is its capacity to support students in expanding vocabulary, generating ideas, and organizing spoken content prior to performance (Du & Daniel, 2024; Liu & Zhao, 2026; Qiao & Zhao, 2023).

Despite these affordances, speaking remains one of the most challenging skills for EFL learners. Many students continue to struggle with limited vocabulary, lack of confidence, and high levels of anxiety when required to speak in English (Horwitz et al., 1986; Yildiz, 2024; Zhang et al., 2024). These difficulties are not solely a reflection of linguistic limitations; rather, they also stem from insufficient preparation before engaging in speaking tasks (Ellis, 2005). Speaking is an interactive and dynamic process that involves the production, reception, and real-time processing of information, requiring learners to manage multiple components simultaneously, including fluency, accuracy, pronunciation, and interactional competence (Brown & Abeywickrama, 2018; Richards, 2008).

In the Indonesian junior high school context, one speaking task that holds particular curricular significance is the monologue in the form of a report text. Report texts are informative in nature and require students to present factual information about a specific topic in a systematic and organized manner (Emilia, 2011; Kemendikbud, 2016). For 9th-grade students, this genre is closely tied to curriculum expectations and commonly involves topics such as animals. However, performing a monologue in this genre is considerably demanding for EFL students, as it requires not only mastery of linguistic elements but also the ability to organize factual information, structure speech coherently, and deliver it communicatively (Burns & Goh, 2012). Brown and Abeywickrama (2018) distinguish the monologue as a form of extended discourse in which the speaker must sustain speech over a longer period without turn-taking, placing greater cognitive demands on the learner. Levelt's (1989) model of speech production and Skehan's (1998) cognitive framework further explain why unassisted monologue tasks are particularly taxing: learners must simultaneously conceptualize content, formulate language, and articulate it fluently under real-time pressure.

One well-established strategy for addressing these cognitive demands is pre-speaking planning. Research by Ellis (2005) and Yuan and Ellis (2003) generally shows that providing learners with planning time before a speaking task can improve fluency and complexity, although findings for accuracy are more mixed. Planning allows speakers to allocate cognitive resources more efficiently, reducing the burden of real-time processing and enabling greater focus on content and language form. In this context, AI tools such as ChatGPT present a promising means of structuring this preparatory phase. By enabling students to interact with an AI system before speaking performance, learners can develop ideas, select appropriate vocabulary, and organize their speech content in a

scaffolded and responsive environment. Monologue report text tasks are particularly suitable for AI-assisted preparation because students must retrieve factual content, organize it in a predictable generic structure, and select topic-specific vocabulary before speaking, all of which an AI tool can help scaffold.

This function of AI as a preparation tool is theoretically grounded in Vygotsky's (1978) concept of the Zone of Proximal Development (ZPD) and scaffolding, the idea that learners can achieve higher levels of performance when supported by a more capable mediator, whether human or technological. AI, in this sense, functions as a form of technological scaffolding, providing just-in-time support that helps learners bridge the gap between their current ability and their potential performance (Prasad & Swaroopa, 2025). This scaffolding function is particularly relevant in EFL contexts where learners may lack access to human interlocutors for meaningful pre-speaking preparation.

Empirical evidence supports the value of AI in spoken language preparation and practice. Yildiz (2024) found that integrating ChatGPT into EFL instruction significantly improved students' speaking self-efficacy, while Du and Daniel (2024), in a systematic review of AI-powered chatbots for EFL speaking practice, found that chatbots can support learners' speaking development. Qiao and Zhao (2023) found in a randomized controlled trial that Duolingo-based instruction led to significantly greater improvement in L2 speaking skills than traditional instruction. Zou et al. (2023) showed that AI-assisted automated feedback could lead to significant gains in speaking skills. More broadly Zhang et al. (2024) found that AI-assisted speaking activities positively influenced learners' enjoyment, reduced anxiety, and increased willingness to communicate.

However, most existing studies have examined AI in the context of interactive speaking practice or post-performance feedback, rather than as a preparation tool for genre-specific monologue tasks (Qiao & Zhao, 2023; Yildiz, 2024; Zou et al., 2023). The specific role of AI in supporting structured, genre-based preparation, particularly for academic monologues such as report texts, remains underexplored. To our knowledge, previous studies have not examined how learners use AI during a short, structured planning phase immediately before a monologue, nor how the quality of students' own prompts relates to their speaking gains.

This gap is significant because the quality of preparation is not only determined by the availability of AI, but also by how students engage with it. In this study, prompt quality refers to how specific a student's prompt is about the topic, genre, and desired support (specificity), how purposefully it directs the AI toward preparation needs (strategic use), and how complex the student's own language input is (linguistic complexity).

Against this backdrop, the present study aims to fill this gap by investigating the effect of AI-assisted preparation on 9th-grade EFL students' monologue speaking performance in a report text task about animals. Unlike prior studies, this research adopts a mixed methods approach that combines pre-experimental quantitative measurement with qualitative analysis of student-generated prompts, thereby offering a more comprehensive understanding of how students interact with AI during the preparation stage. Speaking performance is operationalized as the total score on five rubric components (grammar, vocabulary, comprehension, fluency, and pronunciation), which AI-assisted preparation is expected to support mainly through idea generation, vocabulary selection, and content organization before speaking.

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

The following research questions guide this inquiry: (1) Does AI-assisted preparation significantly improve students' monologue speaking performance? (2) How do patterns in students' AI prompts, in terms of specificity, strategic use, and linguistic complexity, relate to their level of speaking improvement?

2. Method

2.1 Research Design

This study employs a mixed methods design (Creswell & Creswell, 2018), combining a pre-experimental one-group pre-test–post-test design (quantitative) with qualitative content analysis of student AI prompts. The absence of a control group is a recognized limitation of the pre-experimental design; however, it was necessitated by the contextual constraint that only one ninth-grade class was available at the research site. A single class of 20 ninth-grade students received the treatment, and three students with contrasting N-Gain levels were selected for prompt analysis. Both strands were integrated at the interpretation stage.

Table 1. Pre-Experimental One-Group Pre-Test–Post-Test Design

Group	Pre-Test	Treatment	Post-Test
Single Group (9th Grade)	O_1	X	O_2

Note. O_1 = Pre-Test; O_2 = Post-Test; X = Treatment (AI-assisted preparation).

2.2 Participants

Participants were 20 ninth-grade students from a Madrasah Tsanawiyah in Bekasi, selected through purposive sampling. Pre-test scores ranged from 5 to 13 (SMI = 25). For the qualitative strand, three students were selected via maximum variation sampling to represent no improvement (N-Gain = 0.00), low improvement (N-Gain = 0.16), and moderate improvement (N-Gain = 0.54). The three cases were selected to illustrate contrasting levels of improvement rather than to represent the whole class; they are therefore treated as illustrative cases, not as a basis for generalization.

2.3 Instruments

Speaking performance was measured through a pre-test and post-test using the Brown and Abeywickrama (2018) rubric (five components: grammar, vocabulary, comprehension, fluency, and pronunciation; SMI = 25). Prior to data collection, the assessment instruments were validated by three expert validators using Aiken's V coefficient (Aiken, 1985), computed as $V = \Sigma s / [n(c - 1)]$, where $s = r - l_0$ (r is a validator's rating and l_0 is the lowest rating on the scale), n is the number of validators, and c is the number of rating categories. Across 18 indicators, 17 items yielded $V \geq 0.80$, reflecting high content validity. One indicator yielded V = 0.67; however, the overall mean Aiken's V was 0.97, considered high. The full results are presented in Table 2.

Table 2. Aiken's V Content Validity Results

Indicator	V1	V2	V3	Aiken's V	Category
Visual clarity of images with instructions	5	5	5	1.00	High
Suitability of topic selection with 9th grade MTs curriculum	5	5	5	1.00	High
Relevance of animals to 9th grade MTs students	4	4	5	0.83	High
Potential of images to stimulate language production	3	4	4	0.67	Moderate
Choice of words appropriate for 9th grade MTs level	5	5	5	1.00	High
Clarity of grammar rubrics for each level	5	5	5	1.00	High
Clarity of vocabulary rubrics for each level	5	5	5	1.00	High
Clarity of comprehension rubrics for each level	5	5	5	1.00	High
Clarity of fluency rubrics for each level	5	5	5	1.00	High
Clarity of pronunciation rubrics for each level	5	5	5	1.00	High
Suitability of grammar assessment aspects	5	5	5	1.00	High

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

Indicator	V1	V2	V3	Aiken's V	Category
Suitability of vocabulary assessment aspects	5	5	5	1.00	High
Suitability of comprehension assessment aspects	5	5	5	1.00	High
Suitability of fluency assessment aspects	5	5	5	1.00	High
Suitability of pronunciation assessment aspects	5	5	5	1.00	High
Relevance of rubric to monologue report text assignment	5	5	5	1.00	High
Clarity of assessment rubric	5	5	5	1.00	High
Clear language and instructions	5	5	5	1.00	High
Mean Aiken's V (Overall)	—	—	—	0.97	High

Note. V = Aiken's V coefficient; n = 3 validators; scale 1–5. High = V ≥ 0.80; Moderate = V 0.60–0.79 (Aiken, 1985).

2.4 Speaking Task, Assessment, and AI-Assisted Treatment

Students delivered a monologue about an animal based on a topic card, scored by two independent raters. Four picture cards were used (dolphin, eagle, penguin, and elephant), each showing one animal together with a glossary of supporting vocabulary that students could use in both the pre-test and the post-test. The cards were printed and distributed randomly among students, so some students received the same animal while others received different animals. The four cards were designed to be equivalent in difficulty level, and each card contained a glossary with the same number of vocabulary items. Each student used the same card in the pre-test and the post-test. Before the pre-test, students were also given 15 minutes of preparation time, during which they used a printed dictionary and did not use ChatGPT.

The treatment consisted of a structured pre-speaking preparation phase in which students were given 15 minutes to use ChatGPT to develop ideas, select vocabulary, and organize their monologue content before performing the speaking task. All student-submitted prompts were recorded for qualitative analysis.

The AI-assisted treatment followed four steps. First, all students received the same general instruction: to use ChatGPT to prepare a monologue about the animal shown on their topic card. Students were given only this instruction and formulated their own prompts. Second, each student worked individually on their own mobile phone for 15 minutes, using the free version of ChatGPT and no other resources. The session was supervised by the researcher and the class teacher, who did not help students prepare. Third, during the speaking performance, students were not allowed to use ChatGPT, notes, or any other aids apart from the topic card and its supporting vocabulary. Fourth, students' prompts were collected from screenshots of their ChatGPT conversations.

Each monologue was scored by two raters using the Brown and Abeywickrama (2018) rubric. The agreement between the two raters was not calculated statistically; this is acknowledged as a limitation of the study.

2.5 Data Analysis

Quantitative data were analyzed using descriptive statistics, N-Gain (Hake, 1998), the Kolmogorov-Smirnov test, the Shapiro-Wilk test, and a paired samples t-test ($\alpha = 0.05$), all processed using SPSS version 23. N-Gain was used to express each student's improvement relative to the maximum possible improvement on the speaking rubric (SMI = 25), which allows a fairer comparison among students with different pre-test scores. Statistical significance is interpreted only as evidence of a pre-test–post-test difference; because the design has no control group, it is not taken as evidence that AI alone caused the change. Qualitative prompt data were analyzed using directed qualitative content analysis (Hsieh & Shannon, 2005) across three dimensions: (1) specificity, (2) strategic quality, and (3) linguistic complexity of the student's input. Both strands were integrated during the discussion phase.

For the qualitative analysis, each of the three dimensions was operationalized as follows. Specificity referred to how clearly the prompt stated the topic, task, genre, desired structure, and the kind of support requested. Strategic use referred to whether the student directed the AI toward a purposeful preparation need (for example, requesting an outline, vocabulary, or simplified output) and whether follow-up prompts were used. Linguistic complexity referred to the length and sentence structure of the student's own input.

2.6 Integration of Findings

The quantitative and qualitative strands were integrated at the interpretation stage. First, the quantitative results (pre-test and post-test scores and N-Gain) were analyzed to identify the overall pattern and each student's level of improvement. Second, three students with contrasting N-Gain levels were selected and their prompts were analyzed across the three dimensions. Third, the prompt patterns were compared with the N-Gain levels to interpret differences in the degree of improvement; this comparison is presented

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

in Section 3.5 of the Results and discussed in the Discussion. The comparison is descriptive and was not tested statistically.

2.7 Ethical Considerations

Permission to collect data was obtained from the school and the class teacher. Formal ethical clearance was not obtained because no such requirement was in place at the research site at the time of data collection, and written parental consent was not requested. The pre-test and post-test did not affect students' report card grades. Students' identities were protected by reporting them anonymously as Student 1–20, and the prompts they submitted to ChatGPT were recorded for the purpose of the analysis and are reported without identifying information.

3. Results

3.1 Descriptive Statistics

This study involved 20 ninth-grade students who completed a pre-test and post-test of monologue speaking ability in an animal-themed report text task. All performances were assessed using the Brown and Abeywickrama (2018) rubric with a maximum ideal score of 25. The descriptive statistics for pre-test and post-test scores are presented in Table 3.

Table 3. Descriptive Statistics of Pre-Test and Post-Test Scores

	Pre-Test	**Post-Test**
N	20	20
Mean	8.85	12.60
Std. Deviation	2.477	3.662
Minimum	5.00	7.00
Maximum	13.00	19.00

Note. Scores are based on the Brown and Abeywickrama (2018) rubric; maximum ideal score (SMI) = 25.

The pre-test mean score was 8.85 (SD = 2.477), reflecting a low level of initial speaking proficiency. After the AI-assisted preparation, the post-test mean rose to 12.60 (SD = 3.662), representing a mean gain of 3.75 points and suggesting a visible improvement in speaking performance following the intervention.

3.2 N-Gain Analysis

To measure the proportional improvement achieved by each student, the Normalized Gain (N-Gain) formula by Hake (1998) was applied. The N-Gain scores for all 20 students are presented in Table 4.

Table 4. N-Gain Scores Per Student

Student Name	**Pre-Test**	**Post-Test**	**N-Gain**	**Category**
Student 1	9	14	0.31	Moderate
Student 2	6	9	0.16	Low
Student 3	8	13	0.29	Low
Student 4	12	17	0.38	Moderate
Student 5	12	16	0.31	Moderate
Student 6	5	7	0.10	Low
Student 7	7	7	0.00	Low
Student 8	7	11	0.22	Low
Student 9	12	19	0.54	Moderate
Student 10	8	14	0.35	Moderate
Student 11	6	8	0.11	Low
Student 12	8	11	0.18	Low
Student 13	13	19	0.50	Moderate
Student 14	6	8	0.11	Low
Student 15	13	16	0.25	Low
Student 16	8	12	0.24	Low
Student 17	10	14	0.27	Low
Student 18	9	12	0.19	Low
Student 19	8	13	0.29	Low
Student 20	10	12	0.13	Low
Mean	8.85	12.60	0.25	Low

Note. N-Gain categories based on Hake (1998): High ≥ 0.70; Moderate 0.30–0.69; Low < 0.30. SMI = 25.

Table 5. Distribution of N-Gain Categories

Category	**N-Gain Range**	**Number of Students**	**Percentage**
High	≥ 0.70	0	0%
Moderate	0.30 – 0.69	6	30%
Low	< 0.30	14	70%
Mean N-Gain	—	0.25	Low

Note. N-Gain categories adapted from Hake (1998).

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

As shown in Tables 4 and 5, the overall mean N-Gain was 0.25, placing the group in the low improvement category according to Hake's (1998) criteria. Of the 20 students, 14 (70%) were categorized as low-gain and 6 (30%) achieved moderate gain. Only one student (Student 7) showed no improvement, with a pre-test and post-test score of 7 and an N-Gain of 0.00. These findings suggest that although most students' scores increased after AI-assisted preparation, the overall magnitude of improvement was modest.

3.3 Normality Testing and Hypothesis Testing

Before hypothesis testing, the normality of the data was checked using the Kolmogorov-Smirnov test and the Shapiro-Wilk test (Field, 2018). As presented in Table 6, the pre-test results were D = .184 (p = .074) and W = .925 (p = .124). The post-test results were D = .101 (p = .200) and W = .954 (p = .437). In both cases, p > .05, indicating no evidence of violation of the normality assumption. The difference scores (post-test minus pre-test) also showed no evidence of non-normality (Shapiro-Wilk W = .966, p = .665). Accordingly, a paired samples t-test was applied.

Table 6. Normality Test Results (Kolmogorov-Smirnov and Shapiro-Wilk)

Variable	K-S Statistic	K-S p	S-W Statistic	S-W p
Pre-Test	.184	.074	.925	.124
Post-Test	.101	.200	.954	.437

Note. K-S = Kolmogorov-Smirnov; S-W = Shapiro-Wilk; df = 20; significance level α = 0.05.

Table 7. Paired Samples Correlation and T-Test Results

Test	Statistic	df	p	Result
Paired Samples Correlation (PreTest & PostTest)	r = .916	—	< .001	Significant positive correlation
Paired Samples T-Test	t = -9.790	19	< .001	Significant

Note. Significance level α = 0.05. Paired Samples Correlation was computed alongside the t-test using SPSS version 23.

The paired samples t-test yielded t = -9.790 (df = 19, p < .001), with a mean difference of -3.750 (SD = 1.713, SE = .383, 95% CI [-4.552, -2.948]), indicating a highly statistically significant difference between pre-test and post-test scores. The pre-test and post-test scores were also strongly correlated (r = .916, p < .001). Accordingly, the null hypothesis (H_0) is rejected and the alternative hypothesis (H_1) is accepted: students' monologue speaking scores were significantly higher after AI-assisted preparation than before it. Because the design has no control group, this result shows a pre-test–post-test difference and does not by itself show that AI caused the improvement.

Cohen's d for paired samples (d_z) was calculated using the formula d = M_diff / SD_diff (Cohen, 1988), yielding d = 2.19, indicating a large effect size. This indicates a large standardized difference between pre-test and post-test scores; however, given the small sample and the absence of a control group, it should not be read as the isolated effect of the intervention. The contrast between the large effect size (d = 2.19) and the low mean N-Gain (0.25) reflects the different questions the two measures answer, not a contradiction in the data (Hake, 1998; Nissen et al., 2018).

Cohen's d expresses the mean difference relative to the variability of the score differences, so it reflects how consistent the gains were across students. N-Gain expresses each student's gain relative to the maximum gain still available, so it reflects how much of the possible improvement was realized. Here, gains were fairly consistent across students, which produced a large d, while the low pre-test scores relative to the maximum score (SMI = 25) left much room for improvement, which kept the mean N-Gain low. Statistical significance therefore shows that scores differed between the two tests, whereas educational effectiveness depends on the practical size of the gain (a mean gain of 3.75 points on a 25-point rubric and a mean N-Gain of 0.25).

3.4 Qualitative Analysis of Student AI Prompts

To complement the quantitative findings, the AI prompts of three students with contrasting N-Gain levels were analyzed: Student 7 (N-Gain = 0.00, no improvement), Student 2 (N-Gain = 0.16, low improvement), and Student 9 (N-Gain = 0.54, moderate improvement). Prompt analysis focused on three dimensions: (1) specificity of the prompt, (2) strategic use of the AI, and (3) linguistic complexity of the student's input. Because only three students were analyzed, the descriptions below are illustrative patterns and not characteristics of the whole class.

Student 7 (N-Gain = 0.00 – No Improvement)

> ***Student Prompt:*** *"Tell me about elephant."*

Student 7's prompt was notably brief and lacked specificity, direction, or strategic framing. By submitting a single generic imperative without specifying the text genre, desired structure, target audience, or linguistic register, Student 7 failed to leverage the AI as a purposeful preparation tool. This pattern is consistent with Ellis (2005), who argues that unfocused planning produces minimal gains in speaking performance. The absence of any follow-up or iterative prompting suggests that Student 7 engaged minimally with the AI, treating it as a passive information source rather than an interactive preparation partner (Baskara & Mukarto, 2023).

Student 2 (N-Gain = 0.16 – Low Improvement)

> ***Student Prompt:*** *"I want to talk about eagle for my English class. Can you help me with vocabulary and some information about eagle?"*

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

Student 2's prompt demonstrates a higher degree of intentionality compared to Student 7's. The student identified their purpose, named the topic, and requested two specific types of support. However, the prompt remained broad and did not specify the genre (report text), the structural expectations of the monologue, or the particular aspects of the topic they wished to address. Ellis (2005) suggests that focused planning produces greater speaking gains than unfocused planning. Student 2's prompt, purposeful but still broad, may partially explain their low but positive N-Gain of 0.16.

Student 9 (N-Gain = 0.54 – Moderate Improvement)

> ***Student Prompt:*** *"I have a speaking test about report text. My topic is dolphin. Please help me make an outline: classification, description, behavior, and habitat. Also give me some good vocabulary for speaking. Make it simple so I can remember."*

Student 9's prompt reflects the highest level of strategic and genre-aware AI use. The student identified the task type, the genre, the specific topic, desired structural components, the type of support needed, and a practical constraint. This multi-dimensional engagement is consistent with Skehan's (1998) task-focused planning, in which learners allocate cognitive attention to both content organization and linguistic selection. Student 9's prompt essentially directed the AI to function as a structured scaffolding agent, producing output directly aligned with the demands of a report text monologue. The moderate N-Gain of 0.54 was observed alongside this strategic preparation.

The three prompts also differed in the length and structure of the students' own input (linguistic complexity). Student 7's prompt consisted of four words in a single imperative sentence. Student 2's prompt contained 21 words in two sentences, a declarative statement of intent followed by a polite request with two coordinated objects. Student 9's prompt contained 38 words in five sentences, including a colon-introduced list of four outline components and a subordinate purpose clause ("so I can remember"). In these three cases, longer and more structurally varied prompts accompanied higher N-Gain; as with the other two dimensions, this is a descriptive observation and not a tested relationship.

3.5 Integration of Quantitative and Qualitative Findings

Taken together, the quantitative results show that speaking scores were significantly higher after AI-assisted preparation, with modest overall gains (mean N-Gain = 0.25). The three prompt examples show a parallel gradient: the student with the briefest and least specific prompt showed no gain (N-Gain = 0.00), the student with a purposeful but broad prompt showed a low gain (0.16), and the student with a specific, genre-aware prompt showed a moderate gain (0.54). This pattern is suggestive rather than demonstrative: it rests on three cases, was not tested statistically, and does not establish prompt quality as a predictor of improvement.

4. Discussion

The quantitative findings show that students' monologue speaking scores were significantly higher after AI-assisted preparation ($t = -9.790$, $p < .001$). However, the overall mean N-Gain of 0.25 indicates that the magnitude of improvement remained in the low category, with 14 of 20 students (70%) showing low gain, 6 (30%) moderate gain, and none high gain. The qualitative analysis of student AI prompts offers a possible, though tentative, explanation for this finding. Because the study used a one-group pre-test–post-test design without a control group and a small sample ($N = 20$), these results show that students improved after the intervention, not that the intervention caused the improvement; the large effect size ($d = 2.19$) should be read with the same caution. In practical terms, a single brief session of AI-assisted preparation appears to yield modest gains for most students, and simply giving students access to AI may not be sufficient.

The three students analyzed illustrate a gradient in prompt quality that parallels the gradient in their N-Gain scores. This pattern is consistent with the theoretical argument that the effectiveness of AI as a preparation tool may depend on the learner's capacity to engage with it strategically; however, because it rests on only three cases, it should be treated as an emerging pattern that requires confirmation with larger samples, not as a confirmed mechanism. Vygotsky's (1978) concept of scaffolding is instructive here: the AI may function more effectively as a mediating tool when the learner actively directs its support toward specific learning goals.

This finding aligns with the broader literature on pre-task planning in second language acquisition. Ellis (2005) and Yuan and Ellis (2003) suggest that the benefits of planning are maximized when learners attend to both content organization and linguistic form. Effective AI prompting may enable this. Students who used the AI merely as an information retrieval system benefited little, while those who used it for structured content and vocabulary scaffolding showed greater gains in the cases analyzed. This distinction suggests that future AI-assisted speaking preparation should include explicit instruction in prompt formulation. These cases also point to the role of teacher guidance: simply giving students access to AI may not be sufficient, and students may need explicit modelling of how to formulate specific, task-oriented prompts (for example, by naming the genre, the structure, and the kind of language support needed).

The low overall N-Gain may also reflect structural constraints. The 15-minute preparation window is brief, particularly for students using AI as a speaking preparation tool for the first time. Some students may have spent the initial minutes orienting to the interface rather than engaging productively. This was not observed systematically in this study, but it would be consistent with what Baskara and Mukarto (2023) identify as limited AI literacy among EFL learners. Additionally, affective factors should not be overlooked. Horwitz et al. (1986) and Zhang et al. (2024) identify speaking anxiety as a persistent barrier, and its reduction may not materialize fully within a single brief preparation session.

Taken together, the quantitative and qualitative findings converge on a tentative conclusion: students' speaking scores improved significantly after AI-assisted preparation, but the gains were modest overall, and the one-group design and small sample do not allow the improvement to be attributed to AI alone. In the three cases analyzed, more specific, strategic, and task-oriented prompts appeared alongside greater gains, which suggests that how students engage with AI may matter as much as having

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

access to it. These interpretations should be read in light of the limitations discussed below, particularly the absence of a control group and the small sample.

Several limitations of this study need to be acknowledged. First, the single-group pre-experimental design lacks a control group, which limits causal inference. Second, the 15-minute preparation window likely limits the magnitude of improvement, as students unfamiliar with AI tools need time to adapt to the interface. Third, inter-rater reliability coefficients (e.g., Cohen's Kappa or ICC) were not formally calculated; future studies should complement expert validation with quantitative inter-rater agreement indices. Because the two raters' scores were not compared statistically, the objectivity of the speaking scores could not be verified quantitatively. Fourth, each student used the same picture card in both tests, which may have introduced a familiarity effect. Although the four cards were designed to be equivalent in difficulty and glossary size, students spoke about different animals, so small differences in topic familiarity cannot be fully ruled out. Fifth, the small sample size (N = 20) limits statistical power and generalizability. Sixth, the qualitative analysis was limited to the prompts of three students, so the patterns related to prompt quality are illustrative and cannot be generalized to the whole class. The three analyzed students also spoke about different animals (elephant, eagle, and dolphin), so differences in topic familiarity cannot be fully excluded as a factor in the prompts. Seventh, initial proficiency may confound the comparison between prompt quality and gains: the six students with moderate gains had a mean pre-test score of 11.00, compared with 7.93 for the 14 students with low gains, so students who began with higher scores may have both written stronger prompts and gained more. Eighth, because there was no control group, the improvement cannot be separated from the effects of repeated testing, practice, or maturation, even though preparation time was provided in both tests.

5. Conclusion

This study examined the effect of AI-assisted preparation on the monologue speaking performance of 20 ninth-grade EFL students using a mixed methods design. Quantitatively, the results demonstrated a statistically significant improvement in students' speaking scores ($t = -9.790$, $df = 19$, $p < .001$). However, the overall mean N-Gain of 0.25 indicates that the magnitude of gains was modest. Qualitative analysis of three students' prompts suggested that more specific, genre-aware prompts were associated with greater speaking gains, although this pattern is illustrative and requires confirmation with a larger sample. Because the study used a one-group design without a control group, the findings should be interpreted as improvement after the intervention, not as evidence that AI alone caused it.

These findings carry several pedagogical implications. First, the integration of AI into EFL speaking instruction should be accompanied by explicit instruction in prompt formulation. Second, longer or repeated AI-assisted preparation sessions may be worth exploring, since the 15-minute session used here produced modest gains. Third, teachers should attend to affective dimensions of AI use, providing additional support for students with high anxiety or low self-efficacy. Future research should explore the long-term effects of repeated AI-assisted preparation, the role of teacher-guided prompt scaffolding, and the specific linguistic features of student-AI interactions that most strongly predict speaking improvement.

6. References

Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. Educational and Psychological Measurement, 45(1), 131–142.

Baskara, F. X. R., & Mukarto, F. X. (2023). Exploring the implications of ChatGPT for language learning in higher education. Indonesian Journal of English Language Teaching and Applied Linguistics, 7(2), 343–358. https://doi.org/10.21093/ijeltal.v7i2.1387

Brown, D. H., & Abeywickrama, P. (2018). Language Assessment: Principles and Classroom Practices (3rd ed.). Pearson.

Burns, A., & Goh, C. C. M. (2012). Teaching Speaking: A Holistic Approach. Cambridge University Press.

Chen, L., Chen, P., & Lin, Z. (2020). Artificial Intelligence in Education: A Review. IEEE Access, 8, 75264–75278. https://doi.org/10.1109/ACCESS.2020.2988510

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates.

Creswell, J. W., & Creswell, J. D. (2018). Research Design: Qualitative, Quantitative, and Mixed Methods Approaches (5th ed.). SAGE Publications.

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, Article 100230. https://doi.org/10.1016/j.caeai.2024.100230

Ellis, R. (2005). Planning and task-based performance: Theory and research. In R. Ellis (Ed.), Planning and Task Performance in a Second Language (pp. 3–34). John Benjamins.

Emilia, E. (2011). Pendekatan Genre-Based dalam Pengajaran Bahasa Inggris. Rizqi Press.

Field, A. (2018). Discovering Statistics Using IBM SPSS Statistics (5th ed.). SAGE Publications.

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. American Journal of Physics, 66(1), 64–74. https://doi.org/10.1119/1.18809

Haristiani, N. (2019). Artificial Intelligence (AI) Chatbot as Language Learning Medium: An inquiry. Journal of Physics: Conference Series, 1387(1). https://doi.org/10.1088/1742-6596/1387/1/012020

Horwitz, E. K., Horwitz, M. B., & Cope, J. (1986). Foreign Language Classroom Anxiety. The Modern Language Journal, 70(2), 125–132.

Hsieh, H.-F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. Qualitative Health Research, 15(9), 1277–1288. https://doi.org/10.1177/1049732305276687

Huang, X., Zou, D., Cheng, G., Chen, X., & Xie, H. (2023). Trends, Research Issues and Applications of Artificial Intelligence in Language Education. Educational Technology and Society, 26(1), 112–131. https://doi.org/10.30191/ETS.202301_26(1).0009

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

Kemendikbud. (2016). Silabus Mata Pelajaran Sekolah Menengah Pertama/Madrasah Tsanawiyah: Mata Pelajaran Bahasa Inggris. Kementerian Pendidikan dan Kebudayaan.

Levelt, W. J. M. (1989). Speaking: From Intention to Articulation. MIT Press.

Liu, G. L., & Zhao, X. (2026). A scoping review of AI-mediated informal language learning: Mapping out the terrain and identifying future directions. ReCALL, 38(1), 111–130. https://doi.org/10.1017/S0958344025100359

Nissen, J. M., Talbot, R. M., Thompson, A. N., & Van Dusen, B. (2018). Comparison of normalized gain and Cohen's d for analyzing gains on concept inventories. Physical Review Physics Education Research, 14(1), 010115. https://doi.org/10.1103/PhysRevPhysEducRes.14.010115

Prasad, N. B., & Swaroopa, N. (2025). ChatGPT as a scaffolding tool in English Language Teaching. Literary Vibes: An International Journal of English Language and Literature Studies, 14(1). https://doi.org/10.5281/zenodo.14898233

Qiao, H., & Zhao, A. (2023). Artificial intelligence-based language learning: illuminating the impact on speaking skills and self-regulation in Chinese EFL context. Frontiers in Psychology, 14. https://doi.org/10.3389/fpsyg.2023.1255594

Richards, J. C. (2008). Teaching Listening and Speaking: From Theory to Practice. Cambridge University Press.

Skehan, P. (1998). A Cognitive Approach to Language Learning. Oxford University Press.

Vygotsky, L. S. (1978). Mind in Society: The Development of Higher Psychological Processes. Harvard University Press.

Yildiz, C. (2024). ChatGPT Integration in EFL Education: A Path to Enhanced Speaking Self-Efficacy. Novitas-ROYAL, 18(2), 167–182. https://doi.org/10.5281/zenodo.13861137

Yuan, F., & Ellis, R. (2003). The effects of pre-task planning and on-line planning on fluency, complexity and accuracy in L2 monologic oral production. Applied Linguistics, 24(1), 1–27. https://doi.org/10.1093/applin/24.1.1

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? International Journal of Educational Technology in Higher Education, 16(1), Article 39. https://doi.org/10.1186/s41239-019-0171-0

Zhang, C., Meng, Y., & Ma, X. (2024). Artificial intelligence in EFL speaking: Impact on enjoyment, anxiety, and willingness to communicate. System, 121, Article 103259. https://doi.org/10.1016/j.system.2024.103259

Zou, B., Du, Y., Wang, Z., Chen, J., & Zhang, W. (2023). An Investigation Into Artificial Intelligence Speech Evaluation Programs With Automatic Feedback for Developing EFL Learners' Speaking Skills. SAGE Open, 13(3). https://doi.org/10.1177/21582440231193818