



ChatGPT-Assisted Feedback and Students' Grammatical Accuracy in Writing

Dewi Fortuna Nawir¹, Rafi'ah Nur², Ammang Latifa³

^{1,2,3}Ilmu Pendidikan, Universitas Muhammadiyah Parepare, Parepare, Sulawesi Selatan

Article Info	Abstract
Received: 2026-04-07 Revised: 2026-05-19 Accepted: 2026-09-20 Keywords: ChatGPT-assisted feedback; grammatical accuracy; EFL writing; corrective feedback; mixed-methods research DOI: 10.24256/ideas.v14i2.11682 Corresponding Author: Dewi Fortuna Nawir dewifn08@gmail.com Ilmu Pendidikan, Universitas Muhammadiyah Parepare, Parepare, Sulawesi Selatan	<i>enerative artificial intelligence is increasingly integrated into English language learning, yet empirical evidence on its effectiveness for improving grammatical accuracy among Indonesian senior high school EFL learners remains limited; most existing studies have focused on university students or on general perceptions rather than on grammatical accuracy as a specific, measurable outcome. This gap motivated the present study, which investigated the effectiveness of ChatGPT-assisted feedback in improving eleventh-grade students' grammatical accuracy in writing and explored their learning experiences during AI-assisted revision. Grammatical accuracy was selected as the primary outcome because it is one of the most persistent and measurable challenges in EFL writing and underlies clear written communication. An embedded mixed-methods design combined a quasi-experimental pre-test-post-test control-group design with a questionnaire and semi-structured interviews. Sixty-six eleventh-grade students at SMAN 2 Parepare participated; the experimental group (n = 33) received ChatGPT-assisted feedback across four writing-revision sessions, while the control group (n = 33) received conventional classroom feedback. The experimental group significantly outperformed the control group on the independent post-test ($t(64) = 5.69, p < .001$, Cohen's $d = 1.40$), and questionnaire and interview findings indicated that ChatGPT-assisted feedback enhanced grammar awareness, writing improvement, motivation, and revision behaviour, while also encouraging critical evaluation of AI-generated suggestions rather than passive acceptance. These findings suggest that ChatGPT is an effective supplementary formative-feedback tool for developing grammatical accuracy rather</i>

than a replacement for teacher-led instruction. For EFL teachers, the results indicate that structured, teacher-mediated AI feedback can be integrated into writing classes without displacing the teacher's pedagogical role. The study contributes empirical evidence for embedding structured AI-assisted feedback into EFL writing pedagogy at the secondary level, and offers a replicable model for classrooms facing large enrolments and limited time for individualized correction.

1. Introduction

Writing is one of the most essential productive skills in English language learning because it enables learners to communicate ideas, construct arguments, and demonstrate language proficiency in academic contexts. Among the components of effective writing, grammatical accuracy plays a fundamental role in ensuring that written messages are clear, coherent, and comprehensible. Many Indonesian EFL students continue to experience difficulties applying grammatical rules accurately, particularly in tense usage, subject-verb agreement, sentence structure, and word forms, and these persistent problems often reduce the quality of students' writing and their confidence in expressing ideas.

Recent advances in artificial intelligence (AI) have introduced new opportunities to support English language learning. Among these technologies, ChatGPT has attracted considerable attention because it can provide immediate corrective feedback, grammar correction, vocabulary enhancement, idea generation, and writing support. Two distinct sets of considerations, however, accompany this potential.

On one hand, ChatGPT offers clear pedagogical opportunities. Unlike static, delayed teacher feedback, it allows students to revise their writing interactively by receiving explanations, examples, and alternative sentence constructions, and it can therefore reduce the feedback gap that large class sizes typically create. On the other hand, its use raises pedagogical and ethical challenges: students may accept suggestions uncritically, the system can occasionally over-correct or generate contextually inappropriate revisions, and reliance on AI-generated text risks displacing students' own authorship if left unmediated by task design and teacher oversight. Separating these two considerations is necessary because the value of ChatGPT as a feedback tool depends on how the opportunities are harnessed while the risks are actively managed, rather than on the technology alone.

Previous studies have reported that ChatGPT positively influences students' writing performance, grammar development, learning motivation, and engagement, and students generally perceive it as a useful learning assistant. However, most existing studies have examined university students, focused on overall writing quality or learner perceptions, or emphasized general AI adoption rather than grammatical accuracy as a specific learning outcome. Within the

Indonesian secondary-school context specifically, evidence remains sparse. A recent qualitative case study of senior high school students in Indonesia found that ChatGPT supported grammar accuracy, coherence, and autonomous revision, but relied on a small, single-site qualitative sample without an experimental comparison group or a standardized accuracy measure (Haziz, Agustina, Arfah, & Nazri, 2025). This leaves open the question of whether ChatGPT-assisted feedback produces measurable, statistically verifiable gains in grammatical accuracy among senior high school learners when compared against a control condition.

Therefore, a clear research gap remains. Empirical evidence concerning the effectiveness of ChatGPT-assisted feedback in improving grammatical accuracy among Indonesian senior high school EFL learners is still limited, and relatively few studies have combined quantitative evidence of writing improvement with qualitative exploration of students' experiences during AI-assisted revision. Understanding both learning outcomes and learner experiences is important for explaining not only whether ChatGPT is effective but also how students interact with AI-generated feedback throughout the writing process.

The relationships examined in this study can be summarized in a simple conceptual sequence: ChatGPT-assisted feedback is expected to increase students' grammatical noticing (awareness of a mismatch between their own form and a target form); this noticing, supported by explanation and comparison of original and corrected sentences, is expected to translate into revision behaviour; repeated revision across sessions, combined with the motivational effect of immediate and individualized response, is expected to produce measurable gains in independently produced grammatical accuracy. This noticing-revision-accuracy sequence, informed by formative feedback theory (Winstone & Carless, 2020) and Schmidt's (1990) noticing hypothesis, frames both the quantitative hypothesis and the qualitative questions addressed below.

To address the identified gap, and unlike prior studies conducted mainly with university students or without a control condition, the present study employed an embedded mixed-methods design that integrates a quasi-experimental comparison of grammatical-accuracy outcomes with questionnaire and interview evidence of students' experiences, specifically among senior high school learners. This combination is the study's principal point of originality: it isolates grammatical accuracy as a measurable outcome, tests it against a control group, and explains the result through learners' own accounts of noticing, revision, and verification.

Accordingly, this study seeks to answer three research questions: (1) What is the quality of writing produced by eleventh-grade students using ChatGPT-assisted feedback? (2) What behaviours do students demonstrate during writing activities supported by ChatGPT-assisted feedback? and (3) What are students' experiences of using ChatGPT-assisted feedback in English writing? The remainder of this article first describes the quasi-experimental and qualitative procedures used to

answer these questions, then reports the test, questionnaire, and interview findings, and finally discusses their pedagogical implications for EFL writing instruction in secondary schools.

2. Method

Research Design

This study employed an embedded mixed-methods design in which the quantitative strand was primary and the qualitative strand was used to explain and contextualize the quantitative results. The primary component was a quasi-experimental pre-test-post-test control-group design. A true experiment with individual random assignment was not feasible because participants belonged to established school classes and the regular timetable had to be maintained; quasi-experimental designs are appropriate for such settings provided that initial comparability and procedural consistency are examined carefully (Gopalan, Rosinger, & Ahn, 2020). The supporting component consisted of a questionnaire and semi-structured interviews focusing on the experimental group's experience of the intervention.

Integration occurred during interpretation: qualitative themes were compared with the direction and magnitude of the quantitative findings, consistent with the view that mixed-methods research connects numerical trends with contextual explanation (Creswell & Creswell, 2018; Creswell & Plano Clark, 2018).

Table 1. Pre-test–post-test control-group design

Group	Pre-test	Intervention	Post-test
Experimental	O1	X	O2
Control	O1	C	O2

Setting and Participants

The study was conducted at SMAN 2 Parepare with 66 eleventh-grade students from the science stream. Two intact classes served as the sampling units, with 33 students assigned to the experimental condition and 33 to the control condition. Pre-test results indicated close initial comparability: the experimental group's mean was 51.18, and the control group's mean was 50.66, a descriptive difference of only 0.52 points.

The experimental group received ChatGPT-assisted grammatical feedback, whereas the control group received conventional classroom feedback; the comparison was therefore between two feedback environments rather than between practice and no practice. "Both groups were taught by the same teacher, namely the researcher herself, which helped minimize potential instructional bias between the two conditions.

Variables and Operational Definitions

The independent variable was the feedback condition. ChatGPT-assisted feedback referred to a structured process in which students submitted their own

sentences or draft excerpts, requested identification and explanation of grammatical problems, evaluated the suggestions, and revised their texts. Conventional feedback referred to the regular non-ChatGPT support provided during comparable writing activities. The dependent variable was grammatical accuracy in an independently produced narrative text, operationalized through an analytic rubric assessing subject-verb agreement, tense and aspect, articles, prepositions, and sentence or clause construction.

Student experience was treated as a supporting construct. It included perceived ease of use, grammar awareness, confidence, writing improvement, transfer of learning, motivation, autonomy, error reduction, enjoyment, and affective response.

Research Instruments and Instrument Quality

Three instruments were used: a writing test with an analytic grammatical-accuracy rubric, a student-experience questionnaire, and a semi-structured interview protocol. The writing test was administered as both pre-test and post-test; students wrote an individual narrative text of at least 250 words without ChatGPT, dictionaries, or grammar-checking applications, so that the outcome reflected independently demonstrated grammatical control.

The rubric, questionnaire, and interview protocol were informed by the five focal grammatical categories and two supporting mechanics criteria established in the theoretical framework (subject-verb agreement, tense and aspect, articles, prepositions, and clause/sentence structure, plus word choice and punctuation), reflecting content alignment with established EFL grammatical-accuracy constructs. No formal pilot testing, expert-judgment scoring, or reliability coefficient (e.g., Cronbach's alpha) was computed for the rubric, questionnaire, or interview protocol in the present study, which is acknowledged as a limitation below.

Reporting this evidence strengthens confidence that the instruments measured the intended constructs consistently across raters and administrations. Writing-test scoring was conducted using the analytic rubric described above. conducted by a single rater; a second independent rater was not employed, so inter-rater reliability (e.g., Cohen's kappa or ICC) could not be computed, which is acknowledged as a limitation of the present analysis

The questionnaire contained ten statements rated on a five-point Likert scale, each representing one aspect of students' experience: ease of use, grammar awareness, confidence, perceived writing improvement, transfer of learning, motivation, autonomy, error reduction, enjoyment, and affective response. It was administered to the experimental group after the intervention. Mean values of 4.00 or above were interpreted as strongly positive, 3.00-3.99 as moderate, and below 3.00 as negative; this interpretation was used descriptively rather than as a claim that response categories were psychologically equidistant.

Semi-structured interviews were conducted with selected experimental-group students. Ten students ($n = 10$) from the experimental group were purposively selected to represent high, moderate, and low post-test achievement levels. The protocol addressed initial attitudes toward writing, grammatical difficulties, usefulness and clarity of AI feedback, changes in revision behaviour, motivation, perceived improvement, critical reflection, and intention to continue using the tool.

Research Procedure

The study was conducted in six meetings integrated into the regular English writing schedule, each lasting 80 minutes: a pre-test, four treatment sessions, and a post-test. The writing activities during treatment included narrative, descriptive, expository, and argumentative tasks so that students encountered grammatical decisions across more than one rhetorical context. Figure 1 summarizes the overall procedure.

In the first meeting, both groups completed the pre-test under identical conditions. Across the next four meetings, the experimental group followed a consistent writing-and-revision cycle: approximately 35 minutes for drafting, 30 minutes for obtaining and evaluating ChatGPT feedback, and 15 minutes for revising, reflecting, and submitting the text. Structured prompts directed the tool to identify grammatical errors, explain why a form was inaccurate, and suggest possible revisions without rewriting the entire composition; the teacher observed the interactions, clarified difficult points, and intervened when a response was inaccurate, overly broad, or inconsistent with the learning objective.

The control group completed comparable tasks and revision activities using conventional feedback without ChatGPT. In the final meeting, both groups completed a post-test under conditions equivalent to the pre-test, using a different topic to reduce simple memorization while preserving genre and task demands.

Pedagogical Safeguards

Several safeguards were embedded in the instructional procedure: students completed their own drafts before consulting ChatGPT; structured prompts restricted the system's role to identifying, explaining, and suggesting corrections rather than generating the full text; the teacher monitored the interaction and provided clarification; and both formal tests were completed without AI assistance. These safeguards aligned the use of the tool with formative feedback rather than authorship substitution, and encouraged a critical stance in which students compared suggestions with grammatical rules, requested further explanation, and consulted the teacher or classmates when uncertain.

Data Analysis and Integration

Quantitative data were summarized using means, standard deviations, frequencies, and percentages. An independent-samples t-test compared the experimental and control groups' post-test means at an alpha level of .05, supplemented by Cohen's d calculated from the pooled post-test standard deviation and a 95% confidence interval for the mean difference. Questionnaire

responses were analyzed descriptively by calculating the mean for each aspect. Interview responses were transcribed, coded, grouped into categories, and interpreted thematically, following a flexible procedure for identifying patterned meaning within participants' accounts (Braun & Clarke, 2006; Miles, Huberman, & Saldaña, 2020). Integration was conducted through triangulation during the results and discussion stages, retaining both convergence and divergence between quantitative and qualitative evidence.

3. Result

Grammatical-Accuracy Test Results

The pre-test results indicated that the two groups began at closely comparable levels. The experimental group obtained a mean of 51.18 (SD = 8.30), while the control group obtained a mean of 50.66 (SD = 8.13). The difference of 0.52 points was small relative to the variability within each group. Most students in both classes were classified in the poor category, with recurring difficulties in tense consistency, sentence construction, word order, and other grammatical features. The baseline pattern therefore showed that both groups had substantial room for development and that the experimental group did not begin with a meaningful descriptive advantage.

After the four treatment sessions, both groups achieved higher means, but the increase was larger in the experimental group. The experimental group's post-test mean was 74.72 (SD = 7.91), representing a raw increase of 23.54 points from its pre-test mean. The control group's post-test mean was 64.65 (SD = 6.37), representing a raw increase of 13.99 points. The difference between the groups' raw mean gains was therefore 9.55 points. Because the study used independent intact groups and did not report the correlation required for a paired standardized gain, this difference is interpreted descriptively rather than as a formal difference-in-differences test.

The independent-samples t-test showed that the post-test mean difference of 10.07 points was statistically significant, $t(64) = 5.69$, $p < .001$. The 95% confidence interval for the mean difference was approximately [6.54, 13.60], indicating that the observed advantage was not limited to a negligible range. The pooled-standard-deviation effect size was Cohen's $d = 1.40$, which represents a large standardized difference. The null hypothesis of no post-test difference was therefore rejected. Students who participated in ChatGPT-assisted feedback activities demonstrated substantially stronger independent grammatical accuracy than students who received conventional feedback alone.

Table 2. Descriptive statistics and post-test group comparison

Group	N	Pre-test M	Pre-test SD	Post-test M	Post-test SD
Experimental	33	51.18	8.30	74.72	7.91
Control	33	50.66	8.13	64.65	6.37

Category-level results supported the statistical comparison. Twenty-four students in the experimental group (72.73%) reached the good category in the post-test, and no experimental-group students remained in the poor category. In contrast, most students in the control group remained in the fair category. This movement is educationally meaningful because it indicates that improvement was not produced only by a small number of high-scoring students. A large proportion of the experimental class moved into a level associated with more consistent sentence-level control.

The score pattern does not imply that all grammatical problems were eliminated. The experimental post-test standard deviation of 7.91 shows that students continued to vary in performance, and the questionnaire results later indicated only moderate confidence and autonomy. Nevertheless, the combination of a higher mean, a large standardized effect, and category movement suggests that the intervention influenced more than superficial editing. Students completed the post-test without ChatGPT, so the advantage was demonstrated under independent writing conditions.

Questionnaire Results

Questionnaire responses were positive overall, but the pattern varied across aspects. Perceived writing improvement received the highest mean ($M = 4.4$). Grammar awareness, transfer of learning, and error reduction each received a mean of 4.2. Motivation, enjoyment, and affective response each obtained a mean of 4.1. These values were interpreted as strongly positive. Students therefore tended to view ChatGPT-assisted feedback as useful for recognizing grammatical problems, improving their texts, reducing repeated errors, and sustaining engagement during revision.

Table 3. Summary of questionnaire results by aspect

Aspect	Mean	Interpretation
Perceived writing improvement	4.4	Strongly positive
Grammar awareness	4.2	Strongly positive
Transfer of learning	4.2	Strongly positive
Error reduction	4.2	Strongly positive
Motivation	4.1	Strongly positive
Enjoyment	4.1	Strongly positive
Affective response	4.1	Strongly positive
Ease of use	3.9	Moderate
Confidence	3.9	Moderate
Autonomy	3.9	Moderate

Ease of use, confidence, and autonomy each received a mean of 3.9 and were interpreted as moderate. Although these means were close to the threshold for a strongly positive interpretation, their relative position is important. Students did not experience the tool as completely effortless, did not always feel certain about the correctness of their decisions, and did not consider themselves fully

independent. The results challenge the assumption that immediate automated feedback automatically creates confident autonomous writers. Instead, the pattern suggests useful support accompanied by continuing reliance on guidance and verification.

Interview Findings

Interview evidence reinforced the positive questionnaire pattern and clarified what students meant by improvement. Students reported that ChatGPT helped them locate grammatical problems that they had not noticed during drafting. The possibility of comparing an original sentence with a suggested correction made the difference between forms visible. When an explanation was understandable, students could connect the correction to a reason, such as the time frame of a narrative or the need for agreement between a subject and a verb. This experience corresponds with the questionnaire's strong grammar-awareness mean.

A second theme concerned revision efficiency. Students described immediate responses as practical because they did not have to wait until a later lesson to discover whether a sentence was problematic. They could ask a follow-up question, revise the draft, and continue writing. The speed of this cycle appeared to make revision feel like an active part of composition rather than a separate remedial activity. Students also indicated that the feedback increased their willingness to revise, which helps explain the strongly positive motivation, enjoyment, and affective-response ratings.

A third theme concerned critical awareness. Students did not regard every response as unquestionably correct. They recognized that ChatGPT could provide inaccurate, inconsistent, or unsuitable suggestions, particularly when the prompt was unclear or the intended meaning had not been explained. Several students reported checking uncertain responses with the teacher, classmates, or other sources before accepting them. This verification behavior is significant because it shows emerging evaluative judgment. It also explains why confidence and autonomy remained moderate: students benefited from the tool while recognizing that they still needed external confirmation.

A final theme concerned cautious future use. Most interview participants intended to continue using ChatGPT as a supplementary writing tool, especially for locating errors and requesting explanations. At the same time, they expressed concern about becoming too dependent or accepting corrections without learning. Their preferred role for the tool was therefore supportive rather than substitutive. The interviews did not present a simple narrative of enthusiasm; they showed that students could value immediate assistance and still perceive limits.

Integration of Quantitative and Qualitative Findings

The three data sources converged on the conclusion that ChatGPT-assisted feedback supported grammatical development. The large post-test difference showed measurable improvement, the questionnaire indicated strong perceived

writing improvement and grammar awareness, and the interviews described concrete processes of error detection, explanation, comparison, and revision. Because the post-test excluded AI assistance, the quantitative effect was compatible with students' reports that some learning transferred beyond the assisted drafts.

The data also converged on the need for guidance. Moderate confidence and autonomy were consistent with interview reports of uncertainty and verification. Rather than weakening the result, this convergence provides a more realistic interpretation. The intervention was effective not because students became immediately independent of all support, but because a structured environment enabled them to use rapid feedback while retaining access to teacher judgment.

Effect of ChatGPT-Assisted Feedback on Grammatical Accuracy

The principal finding was that students who received ChatGPT-assisted feedback achieved substantially higher post-test grammatical-accuracy scores than students who received conventional feedback. The groups began with nearly identical means, but the experimental group finished 10.07 points higher and demonstrated a large standardized effect. The result supports the quantitative prediction of the conceptual framework and indicates that the intervention produced more than the improvement expected from repeated writing practice alone. Both groups improved, which is reasonable because both completed multiple writing tasks and received feedback. The larger experimental gain suggests that the distinctive features of the AI-assisted cycle—immediacy, individualized attention, explanation, and repeated opportunities for revision—added instructional value.

The finding is consistent with Mahapatra's (2024) mixed-methods intervention, which found that ChatGPT used as formative feedback positively affected ESL students' academic writing. It also aligns with Song and Song's (2023) report that AI-assisted instruction enhanced EFL writing skills and motivation. The present study extends that evidence in three ways. First, it focuses on senior high school rather than university students. Second, it isolates grammatical accuracy as the primary measured outcome. Third, it evaluates independent writing after the tool has been removed. This last point is essential because an AI-assisted draft can be more accurate simply because the system corrected it. A post-test completed without ChatGPT provides stronger evidence that at least part of the learning was available to students during unaided production.

The large effect should nevertheless be interpreted within the boundaries of the design. The classes were intact rather than individually randomized, the intervention lasted four treatment sessions, and the outcome was measured immediately after the treatment. These conditions support a conclusion about the observed classes and period but do not establish permanent acquisition or universal effectiveness. A delayed post-test would be needed to determine whether the advantage remained after several weeks. Moreover, the analytic score combined several grammatical dimensions; the available results do not show

whether tense, agreement, articles, prepositions, or sentence construction improved equally. The result is therefore strong evidence of overall grammatical-accuracy development, not proof that every targeted feature was mastered.

The category movement offers an additional perspective on practical significance. Nearly three quarters of the experimental group reached the good category and none remained in the poor category. This distribution suggests that the intervention was not useful only for students who were already relatively proficient. Immediate feedback may be particularly supportive for learners who hesitate to ask repeated questions in a whole-class setting or who need more examples than a teacher can provide during one lesson. At the same time, variability remained, indicating that AI feedback does not remove differences in prior knowledge, attention, reading comprehension, or willingness to revise.

Why the Intervention May Have Worked

The first explanatory mechanism is grammatical noticing. During drafting, learners often concentrate on meaning and may not detect a form that a reader would recognize as inaccurate. ChatGPT made these discrepancies salient by identifying a problem and placing an alternative next to the original. According to Schmidt (1990), conscious attention to a linguistic feature is an important condition for learning. The interviews indicate that students did not only see corrected sentences; they compared forms and used explanations to understand why a change was recommended. This sequence is particularly relevant for tense consistency and sentence structure, where the correct choice depends on relationships across a larger unit of text.

The second mechanism is the reduction of feedback delay. A teacher may return a text after the learner has shifted attention to another topic. ChatGPT responded while the drafting decision was still recent. Timeliness alone does not guarantee accuracy, but it can strengthen the association between the learner's intention, the produced form, and the explanation. Shute (2008) notes that formative feedback is most useful when it is specific and available in time to guide improvement. In the intervention, students could revise during the same lesson, creating a compact loop of production, feedback, evaluation, and action.

The third mechanism is individualized interaction. In a whole-class lesson, one explanation must often serve many students, even though their errors differ. ChatGPT allowed each learner to ask about the sentence currently causing difficulty. Follow-up questions could request a simpler explanation or another example. This does not mean that the system understood the learner in a human sense, but the conversational interface made the support responsive to the prompt. Kohnke et al. (2023) describe this personalization as an important affordance of ChatGPT for language learning. For teachers, the tool may redistribute routine explanation and create more time to address meaning, genre, and cases in which the automated response is unreliable.

The fourth mechanism is repeated revision. Students did not use the tool only once; they followed a similar feedback cycle in four sessions. Repetition may have helped them recognize error patterns and apply corrections in later tasks. This interpretation is supported by the questionnaire mean for transfer of learning and by the independent post-test. In process-writing terms, development occurred through recurring movement between drafting and redrafting rather than through passive exposure to rules (Hyland, 2019). The structured prompts also required explanation instead of a complete rewrite, increasing the likelihood that students would process the feedback.

A fifth mechanism is affective accessibility. Some learners experience embarrassment when asking a teacher about a basic error or may avoid requesting clarification in front of peers. A conversational system is available without visible social judgment, and students can repeat a question. The strongly positive motivation, enjoyment, and affective-response means suggest that this accessibility made revision less discouraging. However, emotional comfort should not be confused with trustworthiness. A system that feels patient can still produce an inaccurate answer. The strongest instructional design therefore combines low-stakes questioning with explicit verification.

Student Experience, Motivation, and Feedback Literacy

Students' perceptions help explain why the intervention was sustained across the four sessions. Writing improvement received the highest questionnaire mean, followed by grammar awareness, transfer, and error reduction. These dimensions correspond with the measurable post-test outcome, producing convergence between perceived and demonstrated development. Students did not merely report that the application was modern or interesting; they associated it with specific writing functions. The interviews further identified error location, comparison, explanation, and revision efficiency as the features they valued.

Motivation and enjoyment were also strongly positive. This result is consistent with Song and Song (2023), who found that ChatGPT-assisted language learning could improve writing motivation. Immediate feedback may increase motivation because it makes progress visible. Instead of receiving a page covered with unexplained marks, a learner can address one problem, see a revised sentence, and continue. The sense of progress can encourage persistence. Nonetheless, motivation based only on novelty may decrease over time. Long-term integration should therefore emphasize meaningful learning routines rather than the excitement of using a new technology.

The moderate ratings for confidence and autonomy are equally informative. A simplistic account of educational technology might treat these ratings as weaknesses to be removed as quickly as possible. A more careful interpretation is that students developed appropriate caution. They had enough experience to recognize that responses could be inconsistent, and they sought confirmation when necessary. This behavior reflects the early development of feedback literacy and evaluative judgment (Carless & Boud, 2018). Confidence should mean

confidence in a reasoned decision, not confidence that the tool is always correct.

Autonomy also requires clarification. Frequent access to individualized feedback can make learners less dependent on the timing of teacher comments, but it can create a different dependence if students check every sentence automatically. Genuine autonomy involves knowing when support is needed, framing an effective question, evaluating the response, and gradually internalizing the principle. Nicol and Macfarlane-Dick (2006) connect formative feedback with self-regulation when learners use criteria to monitor their work. The intervention moved in this direction by requiring original drafting and evaluation, but four sessions were unlikely to establish full independence. The moderate autonomy mean is therefore plausible and reinforces the need for a longer, staged approach.

Risks, Limitations of AI Output, and Teacher Mediation

Students' concern about inaccurate feedback is supported by the broader literature. ChatGPT can produce fluent corrections and helpful explanations, but it may over-correct, alter meaning, or fail to maintain relationships across sentences (Fang et al., 2023). The risk is amplified in language learning because students may not possess enough knowledge to recognize an error in the explanation. Fluency creates an authority effect: a confident response can appear more reliable than a hesitant but correct human explanation. Teaching should therefore normalize doubt and verification rather than presenting AI as an answer key.

Prompt design is one form of risk control. Requests should specify the target feature, ask for explanations, preserve the student's meaning, and prohibit full rewriting. For example, a prompt can ask the system to list suspected tense and agreement problems in a table with the original sentence, suggested correction, and reason. Students can then mark whether they accept or reject each suggestion. This structure makes the feedback inspectable and limits the temptation to replace the entire text. It also creates evidence that can be discussed with the teacher.

Teacher mediation remains essential because the teacher understands the curriculum, task purpose, learner history, and acceptable level of assistance. ChatGPT cannot determine whether a correction is developmentally appropriate or whether the learner is using the tool in a way that supports the lesson objective.

The teacher can identify when an explanation is misleading, model how to challenge a response, and connect recurring errors to subsequent instruction. In this sense, AI can expand the amount of feedback while the teacher safeguards its educational direction. Academic integrity is another reason to maintain clear boundaries. Barrett and Pack (2023) show that students and teachers may differ in their judgments about acceptable generative AI use in writing. A grammar-focused feedback activity is ethically different from asking the system to generate an assignment, but the boundary must be explicit.

Schools should define permitted uses, require disclosure when appropriate, and design tasks that make the learning process visible. Blanket prohibition may be difficult to enforce and may prevent the development of critical AI literacy, while

unrestricted use can undermine authorship. A guided middle position is more consistent with the intervention and with UNESCO's (2023) human-centered recommendations.

Privacy also requires practical attention. Students should not enter personal identifiers, sensitive experiences, or confidential school information into an external system. Teachers should select age-appropriate procedures and explain that prompts may be processed beyond the immediate classroom. These issues were not measured as outcomes in the present study, but they are necessary conditions for responsible replication. The effectiveness of a tool does not remove the obligation to consider how data are handled.

Pedagogical Implications

The findings support a phased model for classroom integration. In the first phase, the teacher demonstrates the difference between generation and feedback. Students observe how a complete rewrite can hide the learning process, while a focused explanation allows them to make the decision. In the second phase, students use teacher-provided prompts for one or two grammatical categories. In the third phase, they adapt prompts, compare responses, and justify revisions. In the final phase, they revise independently before consulting the tool and identify which errors they can now correct without assistance. This progression treats AI literacy and grammatical development as related goals.

A revision log can strengthen accountability and learning. The log may contain four columns: original sentence, AI suggestion, student decision, and grammatical reason. Students should not be rewarded for accepting the greatest number of corrections. They should be rewarded for accurate reasoning, preservation of meaning, and recognition of questionable advice. The teacher can sample the logs rather than reading every AI conversation, making the process manageable in a large class. Over time, recurring patterns in the logs can inform mini-lessons on tense, agreement, articles, or sentence boundaries.

Assessment should separate assisted practice from independent performance. ChatGPT can be permitted during formative drafting while quizzes, in-class writing, or final tests are completed without it. This distinction allows students to benefit from scaffolding and gives teachers evidence of transfer. The present study used this principle by excluding ChatGPT from the pre-test and post-test. Schools that use only AI-assisted products for assessment may be unable to determine whether a student can apply the grammatical knowledge independently. Teachers should also preserve attention to meaning and voice. Grammar feedback is useful, but writing is not only the production of error-free sentences. Excessive correction can standardize expression or encourage students to choose forms that sound more sophisticated but do not reflect their intended meaning.

Prompts should instruct the system not to change ideas or style unnecessarily, and students should be encouraged to reject changes that distort voice. Barrot (2023) emphasizes that generative AI should support rather than displace second-language writers' agency.

Professional development is needed for teachers as well as students. Teachers require opportunities to test prompts, identify common failure modes, discuss ethical boundaries, and align AI use with curriculum outcomes. The goal is not technical expertise in model architecture. It is pedagogical judgment about when the tool adds value, when human feedback is necessary, and how students can demonstrate learning. Kohnke et al. (2023) similarly argue that ethical and effective use depends on digital competencies among both teachers and learners.

Limitations and Directions for Future Research

Several limitations should guide interpretation. First, the study involved 66 students from two intact classes in one school. The findings may not represent students in other regions, school types, proficiency levels, or digital-access conditions. Replication across multiple schools would improve external validity and permit analysis of contextual differences. Second, class-level assignment rather than individual randomization leaves open the possibility that unmeasured class characteristics contributed to the outcome, even though the pre-test means were closely comparable.

Third, the intervention included four treatment sessions and an immediate post-test. The design demonstrates short-term effectiveness but not long-term retention. Future research should include delayed post-tests and analyze whether students continue to apply the targeted forms without AI assistance. Fourth, the available score data represent overall grammatical accuracy. Error-level analysis could identify whether ChatGPT-assisted feedback is more effective for local forms such as articles and agreement than for global problems involving clause relationships or tense consistency across a text.

Fifth, the questionnaire means describe perceptions but do not provide detailed psychometric evidence in the present report. Future studies should report instrument development, expert review, pilot testing, reliability, and factor structure where appropriate. Sixth, the interviews captured students' accounts but the study did not analyze complete prompt histories or revision logs. Process data could reveal which AI suggestions were accepted, rejected, or modified and whether accurate uptake predicted post-test improvement.

Future research should also compare multiple feedback configurations: teacher-only, ChatGPT-only, and combined teacher-ChatGPT feedback. Such comparisons could determine whether the strongest outcome results from the technology itself or from a hybrid division of labor. Researchers should examine different prompting strategies, levels of teacher mediation, and learner proficiency. Attention to equity is important because students with unstable internet access, limited devices, or lower digital literacy may benefit less from the same intervention.

Finally, research should investigate potential unintended effects. These include dependency, reduced willingness to tolerate uncertainty, homogenization of style, privacy concerns, and inaccurate correction. Positive mean scores should

not lead to the assumption that more AI use is always better. The relevant question is the minimum and most purposeful use that produces learning while preserving agency. Longitudinal designs combining text analysis, screen recordings, stimulated recall, and delayed assessment would provide a fuller account of this balance.

4. Conclusion

This study found that ChatGPT-assisted feedback significantly improved eleventh-grade students' grammatical accuracy in English writing. The experimental and control groups began with nearly equivalent pre-test means, but the experimental group achieved a substantially higher post-test mean. The post-test difference was statistically significant and represented a large standardized effect. Most experimental-group students moved from poor performance to the good category, while none remained in the poor category. Because the post-test was completed without ChatGPT, the result indicates that the intervention supported at least short-term transfer to independent writing.

Student experience data supported and qualified the statistical result. Learners valued immediate error identification, explanation, comparison of original and corrected forms, revision efficiency, motivation, and perceived improvement. At the same time, confidence, ease of use, and autonomy remained moderate. Students recognized that ChatGPT could be inaccurate and reported verifying uncertain suggestions. These findings show that the tool's value did not lie in unquestioned automation.

It lay in the structured opportunity to notice problems, evaluate feedback, and revise while teacher support remained available. ChatGPT should therefore be positioned as a supplementary formative-feedback resource rather than a replacement for teachers or student reasoning. Effective integration requires original student drafting, focused prompts, explanation rather than full rewriting, revision logs or justification, independent assessment, and explicit attention to ethics and privacy. Under these conditions, generative AI can expand access to timely feedback while preserving learner agency.

English teachers are encouraged to introduce ChatGPT through limited, grammar-focused tasks before allowing broader use. Prompts should ask the system to identify and explain errors without rewriting the complete text. Students should compare suggestions with grammar rules and record why a change is accepted, modified, or rejected. Teachers should review recurring patterns and provide direct instruction when the AI response is unclear or inaccurate.

Students should treat ChatGPT as a source of provisional feedback rather than an authority. They should preserve their own ideas, avoid submitting sensitive information, consult teachers when uncertain, and practice revising without the tool so that learning becomes independent. Schools should establish clear guidance concerning acceptable use, disclosure, privacy, data protection, and academic integrity. Future studies should use larger multi-school samples, longer interventions, delayed post-tests, detailed error analysis, documented prompt-

revision histories, and comparisons among teacher-only, AI-only, and hybrid feedback conditions.

5. Acknowledgement

I would like to express my sincere gratitude to myself for the strength, patience, and perseverance throughout the process of conducting and completing this research. I deeply appreciate every struggle, effort, and commitment that enabled me to reach this stage. My deepest gratitude is also addressed to my supervisors, Prof. Dr. Rafi'ah Nur, M.Hum., and Prof. Dr. Ammann Latifa, M.Hum., for their valuable guidance, support, encouragement, and constructive suggestions during the completion of this research.

6. References

- Bani, M., & Masruddin, M. (2021). Development of Android-based harmonic oscillation pocket book for senior high school students. *JOTSE: Journal of Technology and Science Education*, 11(1), 93-103.
- Coupland, J. (Ed.). (2000). *Small talk*. London, England: Longman.
- Denzin, N. K., & Lincoln, Y. S. (Eds.). (2018). *The SAGE handbook of qualitative research* (5th ed.). SAGE Publications.
- Fraser, B. (1999). What are discourse markers? *Journal of Pragmatics*, 31(7), 931–952. doi:10.1016/S0378-2166(98)00101-5
- Furwana, D., Muin, F. R., Zainuddin, A. A., & Mulyani, A. G. (2024). Unlocking the Potential: Exploring the Impact of Online Assessment in English Language Teaching. *IDEAS: Journal on English Language Teaching and Learning, Linguistics and Literature*, 12(1), 653-662.
- Goffman, E. (1981). *Forms of talk*. University of Pennsylvania Press.
- Herring, S. C. (2007). A faceted classification scheme for computer-mediated discourse. *Language@Internet*, 4(1), 1–37.
- Herring, S. C. (2019). The coevolution of computer-mediated communication and computer-mediated discourse analysis: New insights and future directions. In P. Bou-Franch & P. Garcés-Conejos Blitvich (Eds.), *Analyzing digital discourse: New insights and future directions* (pp. 25–67). Palgrave Macmillan. doi:10.1007/978-3-319-92663-6_2
- House, J., & Kádár, D. Z. (2022). Analysing phatic interaction through speech acts – A discussion note. *Acta Linguistica Academica*, 69(3), 381–391. doi:10.1556/2062.2022.00533
- Husnaini, H. (2022). Development of Self Esteem-Oriented Micro Teaching Materials for IAIN Palopo English Education Students. *IDEAS: Journal on English Language Teaching and Learning, Linguistics and Literature*, 10(1), 538-560.
- Jakobson, R. (1960). Closing statement: Linguistics and poetics. In T. A. Sebeok (Ed.), *Style in language* (pp. 350–377). MIT Press.

- Jucker, A. H. (2021). Methodological issues in digital conversation analysis. *Discourse, Context & Media*, 42, 100520. doi:10.1016/j.dcm.2021.100520
- Koivisto, A., Vepsäläinen, H., & Virtanen, M. T. (Eds.). (2023). *Conversation analytic perspectives to digital interaction: Practices, resources, and affordances*. Helsinki, Finland: Finnish Literature Society. doi:10.21435/sflin.22
- Malinowski, B. (1923). The problem of meaning in primitive languages. In C. K. Ogden & I. A. Richards (Eds.), *The meaning of meaning* (pp. 296–336). Kegan Paul.
- Markham, A., & Buchanan, E. (2012). *Ethical decision-making and internet research: Recommendations from the AoIR Ethics Working Committee (Version 2.0)*. Association of Internet Researchers.
- Masruddin, M., & Nasriandi, N. (2022). Lexical and Syntactical Errors Performed by Junior High School Student in Writing Descriptive Text. *IDEAS: Journal on English Language Teaching and Learning, Linguistics and Literature*, 10(1), 1094-1100.
- Meredith, J. (2019). Conversation analysis and online interaction. *Research on Language and Social Interaction*, 52(3), 241–256. doi:10.1080/08351813.2019.1631040
- Meredith, J., Giles, D., & Stommel, W. (Eds.). (2021). *Analysing digital interaction*. Palgrave Macmillan. doi:10.1007/978-3-030-64922-7
- Meredith, J., & Stokoe, E. (2014). Repair: Comparing Facebook 'chat' with spoken interaction. *Discourse & Communication*, 8(2), 181–207. doi:10.1177/1750481313510815
- Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4), 696–735. doi:10.2307/412243
- Sarjanoja, A.-H., Isomursu, M., & Häkkinen, J. (2013). Small talk with Facebook: Phatic communication in social media. In *AcademicMindTrek '13: Proceedings of the International Conference on Making Sense of Converging Media* (pp. 118–121). ACM. doi:10.1145/2523429.2523449
- Schegloff, E. A. (2007). *Sequence organization in interaction: A primer in conversation analysis (Vol. 1)*. Cambridge University Press. doi:10.1017/CBO9780511791208
- Schiffrin, D. (1987). *Discourse markers*. Cambridge University Press.
- Tagg, C. (2015). *Exploring digital communication: Language in action*. Routledge.
- Walther, J. B. (1996). Computer-mediated communication: Impersonal, interpersonal, and hyperpersonal interaction. *Communication Research*, 23(1), 3–43. doi:10.1177/009365096023001001