INTEGRATING GENERATIVE AI INTO EFL INSTRUCTION: EFFECTS ON STUDENTS' LANGUAGE PERFORMANCE AND LEARNING ENGAGEMENT

Puput Arianto[1] Furqon Edi Wibowo[1], Ernadewi Kartikasari[1]
[1] *Department of English Language Education, Faculty of Cultures and Languages, UIN Raden Mas Said Surakarta, Indonesia*
Corresponding author: puput.ariyanto@staff.uinsaid.ac.id

Abstract

The rapid adoption of generative artificial intelligence (GenAI) in English as a Foreign Language (EFL) instruction has created new opportunities for personalized support, immediate feedback, and interactive practice, yet empirical evidence on its structured, teacher-guided integration and its simultaneous effects on language performance and multidimensional learning engagement remains limited in higher education. This study examined the effects of teacher-guided, GenAI-supported EFL instruction on undergraduate students' language performance and learning engagement. A quasi-experimental pre-test-post-test control group design was used with 68 undergraduate EFL students at an Indonesian university, assigned to an experimental group (n = 34) and a control group (n = 34). Over eight weeks, the experimental group completed teacher-guided GenAI-supported activities covering brainstorming, language practice, feedback, revision, and reflection, while the control group received conventional teacher-led instruction using equivalent materials and time. Language performance and learning engagement were measured with a validated writing test and questionnaire, and data were analyzed using descriptive statistics, paired-samples t-tests, and analysis of covariance (ANCOVA). The experimental group's post-test performance (M = 82.16, SD = 6.74) exceeded the control group's (M = 74.53, SD = 7.02), with a significant effect of instructional condition after controlling for pre-test scores, $F(1, 65) = 18.74$, $p < .001$, partial eta-squared = .224. The experimental group also showed significantly greater gains in cognitive and behavioral engagement. The findings indicate that GenAI supports EFL learning most effectively when embedded in structured, teacher-guided instruction rather than used independently, underscoring the importance of pedagogical scaffolding and critical evaluation of AI-generated content.

Keywords: generative artificial intelligence; EFL instruction; language performance; learning engagement; higher education

INTRODUCTION

The rapid advancement of artificial intelligence technologies, particularly Large Language Models (LLMs) and Generative AI (GenAI) platforms such as ChatGPT, Claude, and Gemini, has initiated a profound paradigm shift in higher education pedagogy (Kasneci et al., 2023; Zhai, 2022). In the domain of English as a Foreign Language (EFL) instruction, GenAI tools have transitioned from novel technological artifacts to integral components of digital learning environments (Kohnke et al., 2023). These platforms offer real-time text generation, automated writing evaluation, conversational practice, and tailored linguistic scaffolding (Hong, 2023; Su et al., 2023). However, while the integration of GenAI offers unprecedented access to interactive target language exposure, higher education institutions face significant challenges in embedding these tools

effectively into structured curricula without diminishing critical thinking or inducing student overreliance (Grassini, 2023; Selwyn, 2019).

Current scholarship in computer-assisted language learning (CALL) highlights the potential of automated systems to foster personalized learning pathways and lower affective barriers (Godwin-Jones, 2022; Jeon, 2024). Scholars observe that GenAI can serve as an adaptive conversational partner and automated feedback generator, enabling EFL learners to obtain immediate, formative feedback on grammar, stylistic nuances, and contextual coherence (Su et al., 2023; Yan, 2023). Despite these theoretical affordances, empirical research on the instructional integration of GenAI presents mixed findings. Unstructured or unguided use of AI tools often leads to passive consumption, superficial revision strategies, and ethical concerns regarding academic integrity (Chiu, 2023; Cotton et al., 2023). Conversely, when AI tools are deployed without explicit pedagogical scaffolding, L2 learners frequently struggle to critically evaluate the accuracy and stylistic appropriateness of AI-generated feedback (Kohnke et al., 2023; Pokrivcakova, 2019).

While existing literature has extensively examined student perceptions, acceptance of AI applications, or isolated automated written corrective feedback (AWCF) mechanics (Hong, 2023; Su et al., 2023), significant empirical gaps persist. First, empirical evidence regarding the impact of GenAI on combined linguistic performance and multidimensional learning engagement in structured university contexts remains scarce (Jeon, 2024; Yan, 2023). Second, much of the prevailing research relies on short-term experimental trials or self-reported survey data, lacking rigorous quasi-experimental control designs that track performance across extended instructional periods (Godwin-Jones, 2022; Grassini, 2023). Third, there is a need to understand how teacher-mediated pedagogical frameworks transform raw GenAI capabilities into structured cognitive and behavioral engagement in non-Western tertiary contexts such as Indonesia (Chiu, 2023; Kasneci et al., 2023).

To address these empirical gaps, this study investigates the impact of an eight-week, teacher-guided GenAI-supported EFL instructional framework on undergraduate students' language performance and learning engagement. Specifically, the study compares structured, AI-scaffolded instruction with conventional teacher-led pedagogy at an Indonesian university. The research addresses the following guiding questions: (1)

Is there a statistically significant difference in post-intervention language performance between students who receive teacher-guided GenAI-supported instruction and those who receive conventional instruction, controlling for pre-intervention performance? (2) How does teacher-guided GenAI-supported instruction influence students' multidimensional learning engagement across behavioral, cognitive, and affective domains?

LITERATURE REVIEW

Theoretical Framework: Sociocultural Theory and Self-Determination Theory

This study is grounded in Vygotsky's (1978) Sociocultural Theory of Learning and Ryan and Deci's (2000) Self-Determination Theory (SDT). Sociocultural theory posits that cognitive development and L2 acquisition occur through mediated interactions within the Zone of Proximal Development (ZPD), wherein more knowledgeable others (MKOs) provide structured scaffolding (Lantolf et al., 2018). In an AI-enhanced classroom, GenAI functions as a digital MKO, offering dynamic, context-sensitive linguistic prompts and feedback (Kohnke et al., 2023; Su et al., 2023). However, human teacher mediation remains essential to scaffold the interaction, guide critical evaluation, and prevent cognitive overload (Grassini, 2023; Zhai, 2022).

Complementarily, SDT asserts that optimal learning engagement and intrinsic motivation depend on satisfying three basic psychological needs: autonomy, competence, and relatedness (Ryan & Deci, 2000). GenAI integration supports student autonomy by allowing learners to control the pacing, depth, and direction of their inquiry during language tasks (Chiu, 2023). Competence is fostered through immediate, non-judgmental corrective feedback that enables iterative skill refinement (Hong, 2023; Yan, 2023). When integrated into a teacher-guided classroom structure, relatedness is sustained through peer collaboration, teacher facilitation, and shared reflective discussion (Jeon, 2024; Kasneci et al., 2023).

GenAI in EFL Pedagogy and Language Performance

Recent empirical syntheses reveal that LLM-driven tools enhance foreign language production by delivering personalized linguistic input and actionable feedback (Godwin-Jones, 2022; Su et al., 2023). Studies on automated written corrective feedback demonstrate that GenAI platforms outperform traditional rule-based grammar checkers

by offering contextualized rephrasing, lexical variety, and discourse organization suggestions (Hong, 2023; Yan, 2023). In higher education settings, interactive LLM prompts encourage learners to notice linguistic gaps between their output and target-like structures (Kohnke et al., 2023). However, researchers emphasize that performance gains are highly contingent on the instructional model. Without explicit teacher guidance on prompt engineering, error identification, and recursive revision, students risk accepting AI-generated hallucinations or grammatically fluent but pragmatically inappropriate language (Cotton et al., 2023; Grassini, 2023).

Multidimensional Learning Engagement in AI-Supported Learning

Learning engagement is a multidimensional construct comprising behavioral (persistence, effort, active participation), cognitive (deep learning strategies, critical thinking, self-regulation), and affective/emotional (interest, enthusiasm, belonging) dimensions (Fredricks et al., 2004). In digital language learning environments, technology can enhance or impede these engagement dimensions depending on task design (Chiu, 2023; Jeon, 2024). Studies indicate that while GenAI tools readily boost initial emotional engagement through novelty and interactive conversational interfaces, sustaining behavioral and cognitive engagement requires deliberate instructional task design (Kasneci et al., 2023; Selwyn, 2019). Cognitive engagement increases when students evaluate AI outputs against target criteria, refine prompts, and justify their editorial choices (Kohnke et al., 2023). Behavioral engagement is supported through iterative drafting cycles and structured peer-teacher feedback loops (Yan, 2023).

Conceptual Model

Synthesizing these theoretical perspectives, the conceptual framework of this study models teacher-guided GenAI instruction as an integrated pedagogical system. In this model, GenAI tools are embedded within a structured, five-stage instructional cycle: brainstorming, target language practice, automated feedback generation, critical revision, and reflective evaluation. This scaffolded cycle acts directly on students' cognitive processing and self-regulation, hypothesized to lead to improvements in both objective language performance and multidimensional learning engagement compared to conventional teacher-led instruction.

RESEARCH METHOD

Research Design

This study employed a quasi-experimental pre-test-post-test control group design (Creswell & Creswell, 2018) conducted over an eight-week instructional period during the 2025/2026 academic year. A quasi-experimental approach was chosen due to the non-random assignment of intact university classes into instructional conditions. The design framework is illustrated in Table 1.

Table 1. *Quasi-Experimental Research Design*

Group	Pre-Test	Intervention (8 Weeks)	Post-Test
Experimental (n = 34)	O1 (Performance Test) O2 (Engagement Survey)	Teacher-Guided GenAI-Supported EFL Instruction	O3 (Performance Test) O4 (Engagement Survey)
Control (n = 34)	O1 (Performance Test) O2 (Engagement Survey)	Conventional Teacher-Led EFL Instruction	O3 (Performance Test) O4 (Engagement Survey)

Participants and Setting

The study was conducted at UIN Raden Mas Said Surakarta, Indonesia. Participants were 68 second-year undergraduate EFL students enrolled in two parallel sections of an Academic Writing and Oral Communication course. Intact classes were assigned to either the experimental group (n = 34; 12 males, 22 females; mean age = 19.8 years) or the control group (n = 34; 10 males, 24 females; mean age = 19.6 years). All participants shared similar socio-cultural backgrounds, native language (Indonesian), and institutional context. Prior to the intervention, independent samples t-tests confirmed no statistically significant baseline differences between the two groups in academic background or general English proficiency. Ethical approval was secured from the institutional review board, and written informed consent was obtained from all participants prior to data collection.

Instructional Interventions

Both groups received two hours of weekly instruction over eight consecutive weeks, covering identical syllabus modules focused on academic essay structure, argumentative discourse, formal register, and oral synthesis. The same course instructor delivered lessons to both groups to eliminate instructor variability. The experimental group's curriculum integrated GenAI platforms (ChatGPT and Claude) through a structured five-

stage pedagogical cycle: (1) brainstorming and ideation, in which students used GenAI to construct topic outlines, explore opposing viewpoints, and expand domain vocabulary; (2) target language practice, in which students drafted academic paragraphs using GenAI-suggested lexical frameworks; (3) automated feedback, in which students submitted drafts to GenAI to request specific feedback on grammar, cohesion, and academic tone; (4) critical revision, in which students systematically verified AI suggestions, rejected hallucinated or inappropriate recommendations, and revised their work; and (5) reflective learning, in which students maintained a reflective log documenting how GenAI feedback altered their linguistic choices. The control group covered identical topics, readings, and writing assignments through traditional teacher presentations, textbook exercises, manual peer-editing workshops, and direct teacher corrections, without student access to AI tools during class hours.

Data Collection Instruments

Language performance was measured using a validated academic writing and synthesis test developed in accordance with CEFR B2 descriptors. The test evaluated four domain criteria: content development (25%), organization and coherence (25%), vocabulary and style (25%), and grammatical accuracy (25%). Pre- and post-tests used equivalent prompt forms, counterbalanced across groups. Test submissions were anonymized and independently scored by two trained raters using a standardized rubric. Inter-rater reliability was high, with an intraclass correlation coefficient (ICC) of .89 for the pre-test and .92 for the post-test.

Learning engagement was measured using an adapted version of the Student Engagement in Learning Scale (Fredricks et al., 2004), modified for EFL contexts. The instrument comprised 21 items on a 5-point Likert scale (1 = strongly disagree to 5 = strongly agree), measuring three subscales: behavioral engagement (7 items; e.g., “I persistently revise my English assignments based on feedback”), cognitive engagement (7 items; e.g., “I critically analyze whether language corrections improve my writing”), and affective/emotional engagement (7 items; e.g., “I feel confident and enthusiastic when working on English tasks”). Construct validity was verified via confirmatory factor analysis, and internal consistency for all subscales exceeded Cronbach's alpha = .84.

Data Analysis

Data were analyzed using SPSS Version 28. Descriptive statistics (means and standard deviations) were calculated for all pre- and post-intervention measures. Paired-samples t-tests assessed within-group performance gains from pre-test to post-test. Analysis of covariance (ANCOVA) evaluated post-intervention group differences in language performance and multidimensional engagement, using pre-test scores as covariates to adjust for minor baseline variations. Assumptions for ANCOVA, including normality (Shapiro-Wilk test), homogeneity of variances (Levene's test), and homogeneity of regression slopes, were tested and satisfied ($p > .05$). Effect sizes were calculated using partial eta-squared, with values of .01, .06, and .14 representing small, medium, and large effects, respectively (Cohen, 1988).

RESULTS AND DISCUSSION

Effects of GenAI-Supported Instruction on Language Performance

Descriptive analysis revealed that both groups improved their language performance scores over the eight-week period. However, the experimental group exhibited a substantially higher mean post-test score compared to the control group. Table 2 summarizes the descriptive results for language performance.

Table 2. *Descriptive Statistics for Language Performance Test Scores*

Group	n	Pre-Test Mean (SD)	Post-Test Mean (SD)	Mean Gain
Experimental	34	65.42 (6.18)	82.16 (6.74)	+16.74
Control	34	64.87 (6.31)	74.53 (7.02)	+9.66

Paired-samples t-tests confirmed that post-test improvements were statistically significant within both the experimental group, $t(33) = 14.82$, $p < .001$, $d = 2.54$, and the control group, $t(33) = 8.12$, $p < .001$, $d = 1.39$.

To test whether the difference in post-intervention language performance between groups was statistically significant, a one-way ANCOVA was conducted, controlling for pre-test performance. The results are presented in Table 3.

Table 3. *ANCOVA Results for Post-Test Language Performance*

Source	Type III SS	df	Mean Square	F	p	Partial η^2
Corrected Model	1258.42	2	629.21	14.18	< .001	.304
Intercept	842.15	1	842.15	18.98	< .001	.226

Pre-Test (Covariate)	274.11	1	274.11	6.18	.015	.087
Group (Condition)	831.64	1	831.64	18.74	< .001	.224
Error	2884.21	65	44.37			
Total	418290.00	68				

After adjusting for pre-test scores, a statistically significant main effect was found for instructional condition on post-test language performance, $F(1, 65) = 18.74$, $p < .001$, with a large effect size (partial eta-squared = .224). The adjusted post-test mean for the experimental group was 82.04 (SE = 1.14), significantly higher than the adjusted post-test mean for the control group (74.65, SE = 1.14).

Effects of GenAI-Supported Instruction on Learning Engagement

Analysis of the multidimensional learning engagement questionnaire revealed post-intervention gains across all three domains (behavioral, cognitive, and affective) for the experimental group. Table 4 reports the baseline and post-intervention engagement scores across both conditions.

Table 4. *Pre- and Post-Intervention Learning Engagement Scores by Subscale*

Engagement Subscale	Group	Pre-Test M (SD)	Post-Test M (SD)	Adjusted F(1,65)	p	Partial η^2
Behavioral	Experimental Control	3.32 (0.48) 3.28 (0.51)	4.38 (0.42) 3.65 (0.46)	38.45	< .001	.372
Cognitive	Experimental Control	3.21 (0.52) 3.25 (0.49)	4.41 (0.39) 3.58 (0.44)	46.12	< .001	.415
Affective	Experimental Control	3.45 (0.50) 3.41 (0.53)	4.12 (0.45) 3.72 (0.48)	12.86	< .001	.165
Overall Engagement	Experimental Control	3.33 (0.41) 3.31 (0.43)	4.30 (0.35) 3.65 (0.38)	52.34	< .001	.446

ANCOVA models controlling for baseline engagement confirmed that the experimental group achieved significantly greater improvements across all engagement dimensions. The largest effect sizes were observed in cognitive engagement (partial eta-squared = .415) and behavioral engagement (partial eta-squared = .372), followed by affective engagement (partial eta-squared = .165).

Discussion

The findings demonstrate that integrating generative AI into structured, teacher-guided EFL instruction significantly enhances undergraduate language performance and multidimensional learning engagement compared to conventional teacher-led methods.

The observed increase in language performance ($F(1, 65) = 18.74$, $p < .001$, partial eta-squared = .224) supports Sociocultural Theory, confirming that GenAI can act as an effective digital MKO when embedded within a mediated instructional framework (Kohnke et al., 2023; Lantolf et al., 2018). While both groups improved over time, the experimental group's superior gains align with prior findings on automated writing evaluation and interactive AI feedback (Hong, 2023; Su et al., 2023). Immediate, non-judgmental feedback allowed experimental students to engage in multiple iterative drafting cycles, reinforcing structural, lexical, and grammatical conventions in real time (Yan, 2023). Crucially, the intervention's reliance on critical evaluation, where students actively evaluated and verified AI suggestions, prevented common pitfalls associated with unguided AI adoption, such as uncritical acceptance of inaccurate outputs (Chiu, 2023; Grassini, 2023).

Regarding learning engagement, the substantial gains in cognitive (partial eta-squared = .415) and behavioral engagement (partial eta-squared = .372) align with Self-Determination Theory (Ryan & Deci, 2000). The structured, five-stage pedagogical cycle satisfied learners' autonomy and competence needs by giving them agency over their drafting processes while providing immediate feedback loops (Chiu, 2023; Hong, 2023). Students maintained high behavioral engagement through active prompt generation and systematic text revision, rather than passively receiving corrections (Kasneci et al., 2023). The significant increase in cognitive engagement further indicates that teacher-scaffolded activities required students to process AI outputs deeply, compare alternative phrasing, and justify editorial decisions (Kohnke et al., 2023; Zhai, 2022). Although affective engagement also improved significantly (partial eta-squared = .165), its smaller effect size relative to cognitive and behavioral dimensions suggests that while AI tools spark initial enthusiasm, long-term learning outcomes rely primarily on sustained cognitive effort and structured task design (Godwin-Jones, 2022; Grassini, 2023).

CONCLUSIONS

This study investigated the effects of teacher-guided GenAI-supported instruction on undergraduate EFL students' language performance and learning engagement. The findings confirm that integrating GenAI platforms into a scaffolded, five-stage

pedagogical cycle yields statistically significant improvements in language performance ($F(1, 65) = 18.74$, $p < .001$, partial eta-squared = .224) and multidimensional engagement compared to conventional instruction. Gains were particularly pronounced in cognitive and behavioral engagement, underscoring the role of active revision and critical evaluation in AI-assisted learning environments.

The theoretical implication of this research lies in extending sociocultural and self-determination frameworks to AI-mediated L2 learning. The results show that GenAI tools operate most effectively not as autonomous substitutes for instruction, but as digital MKOs embedded within teacher-mediated learning contexts. Practically, the study offers a reproducible pedagogical model for tertiary EFL educators, demonstrating how prompt engineering, iterative drafting, and reflective logs can be combined to foster L2 proficiency while safeguarding against passive overreliance.

Limitations and Future Directions

Several limitations should be noted. First, the sample was limited to 68 undergraduate students from a single Indonesian university, which may limit the generalizability of the findings to other cultural or institutional contexts. Second, the eight-week intervention period focused primarily on academic writing and synthesis skills; future research should evaluate the long-term retention of language gains and examine oral communication outcomes over extended academic periods. Finally, future studies could employ qualitative methods, such as think-aloud protocols and semi-structured interviews, to capture deeper insights into students' cognitive prompt-engineering strategies and critical decision-making processes when interacting with GenAI tools.

REFERENCES

Chiu, T. K. F. (2023). The impact of Generative AI (GenAI) on practices, policies and research direction in education: A case of ChatGPT and Midjourney. *Interactive Learning Environments, 31*, 6187–6203. https://doi.org/10.1080/10494820.2023.2253861

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

Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International, 61*(2), 228–239. https://doi.org/10.1080/14703297.2023.2190148

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). SAGE Publications.

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research, 74*(1), 59–109. https://doi.org/10.3102/00346543074001059

Godwin-Jones, R. (2022). Partnering with AI: Intelligent writing assistance and instructed language learning. *Language Learning & Technology, 26*(2), 5–24. https://hdl.handle.net/10125/73474

Grassini, S. (2023). Shaping the future of education: Exploring the potential and consequences of AI and ChatGPT in educational settings. *Education Sciences, 13*(7), 692. https://doi.org/10.3390/educsci13070692

Hong, W. C. H. (2023). The impact of ChatGPT on foreign language teaching and learning: Opportunities in education and research. *Journal of Educational Technology and Innovation, 5*(1), 37–45. https://doi.org/10.61414/jeti.v5i1.103

Jeon, J. (2024). Exploring AI chatbot affordances in the EFL classroom: Young learners' experiences and perspectives. *Computer Assisted Language Learning, 37*(1–2), 1–26. https://doi.org/10.1080/09588221.2021.2021241

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., Kuhn, J., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. *RELC Journal, 54*(2), 537–550. https://doi.org/10.1177/00336882231162868

Lantolf, J. P., Thorne, S. L., & Poehner, M. E. (2018). Sociocultural theory and L2 development. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (3rd ed., pp. 223–247). Routledge.

Pokrivcakova, S. (2019). Preparing teachers for the application of AI-powered technologies in foreign language education. *Journal of Language and Cultural Education, 7*(3), 135–153. https://doi.org/10.2478/jolace-2019-0025

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55*(1), 68–78. https://doi.org/10.1037/0003-066X.55.1.68

Selwyn, N. (2019). *Should robots replace teachers? AI and the future of education.* Polity Press.

Su, Y., Lin, Y., & Lai, C. (2023). Collaborating with ChatGPT in argumentative writing classrooms. *Assessing Writing, 57*, 100752. https://doi.org/10.1016/j.asw.2023.100752

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes.* Harvard University Press.

Yan, D. (2023). Impact of ChatGPT on learners in a L2 writing practicum: An exploratory investigation. *Education and Information Technologies, 28*(11), 13943–13967. https://doi.org/10.1007/s10639-023-11742-4

Zhai, X. (2022). ChatGPT user experience: Implications for education. *SSRN Electronic Journal*, 1–18. https://doi.org/10.2139/ssrn.4312418