Castledown

OPEN ACCESS

Language Education & Assessment
ISSN 2209-3591
https://www.castledown.com/journals/lea/

Language Education & Assessment, 9, 104947 (2026)
https://doi.org/10.29140/lea.2026.104947

ChatGPT Use Patterns and Higher-Order Thinking Performance in EFL Argumentative Writing

WINIA WAZIANA[a] (Corresponding Author)
WIDI ANDEWI[b]
RICCO HERDIYAN SAPUTRA[c]
MUHAMAD MUSLIHUDIN[d]
IRENE BRAINNITA OKTARIN[e]

[a]*Department of Information System, Institut Bakti Nusantara, Indonesia*
winiawaziana@gmail.com

[b]*Department of Information System, Institut Bakti Nusantara, Indonesia*
widiandewi.91@gmail.com

[c]*Department of Information System, Institut Bakti Nusantara, Indonesia*
saputraherdiyanricco@gmail.com

[d]*Department of Information System, Institut Bakti Nusantara, Indonesia*
mmuslihudin415@gmail.com

[e]*Department of Management, STIE Gentiaras, Indonesia*
irenebrain04@gmail.com

Abstract

The growing use of ChatGPT in higher education raises questions about its relationship with students' higher-order thinking (HOT) and how students engage with AI during learning. This study examined whether structured ChatGPT-supported instruction was associated with EFL students' HOT performance in argumentative writing and their patterns of ChatGPT use during the intervention. A quasi-experimental mixed-methods design involved 100 fourth-semester undergraduates drawn from two intact classes, comprising an experimental group (n = 50) and a control group (n = 50). Over 12 weeks, the experimental group completed structured ChatGPT-supported argumentative tasks, while the control group completed comparable tasks through conventional instruction. HOT performance was assessed through supervised pretest and posttest argumentative essays measuring analysis, evaluation, and creation. ANCOVA showed that the experimental group achieved significantly higher adjusted posttest HOT scores than the control group, with creation showing the largest dimension-level difference. Interaction logs showed that reflective-evaluative use was more common

Data Availability Statement: *All relevant data are within this paper.*

than product-oriented use during the supervised ChatGPT-supported tasks. Students frequently evaluated AI suggestions, compared alternatives, and revised selectively. These patterns reflected students' use of the prompting, evaluation, and revision practices introduced during the intervention and were interpreted in relation to cognitive amplification and cognitive offloading. The findings highlight the value of structured AI-supported tasks that maintain students' responsibility for evaluation and revision.

Keywords: ChatGPT, EFL argumentative writing, higher-order thinking, reflective-evaluative AI use, cognitive amplification, cognitive offloading.

Introduction

The expanding use of Generative AI (GenAI) has altered how university students access information, develop ideas, and revise academic work. Among the available GenAI tools, ChatGPT is particularly relevant to language education because it enables learners to interact with an AI system through natural-language prompts and receive immediate, context-sensitive responses (Andewi et al., 2025; Hastomo et al., 2026). In EFL learning, this dialogic capability can support brainstorming, vocabulary exploration, grammar clarification, feedback-seeking, and iterative revision. However, the educational significance of ChatGPT extends beyond its ability to improve linguistic accuracy. The more consequential question is whether students use the tool to deepen their reasoning or merely to complete academic tasks more efficiently.

The distinction between cognitive support and cognitive substitution is particularly consequential in EFL argumentative writing, where effective performance extends beyond grammatical accuracy. Students must identify the central issue, distinguish relevant information, examine relationships among ideas, evaluate evidence, address counterarguments, and construct a coherent position. These processes correspond to HOT, particularly the dimensions of analysis, evaluation, and creation. Research on EFL writing has shown that critical-thinking-oriented instruction can improve students' ability to evaluate claims and develop arguments (Li & Liu, 2021; Lu & Xie, 2019). Emerging studies further suggest that ChatGPT can support academic writing, argumentation, and critical engagement when it is incorporated into structured learning activities rather than used solely as a source of ready-made answers (Darmawansah et al., 2025; Liang & Wu, 2024; Mahapatra, 2024).

Nevertheless, the use of ChatGPT creates a pedagogical tension. The same tool that offers alternative viewpoints, counterarguments, and feedback may also reduce the cognitive effort required to complete a task. Cognitive offloading occurs when individuals transfer part of the cognitive demands of a task to an external resource (Risko & Gilbert, 2016). In AI-supported learning, offloading may become pedagogically problematic when students request complete answers, accept AI-generated arguments without evaluation, or reproduce generated text with minimal revision. Conversely, cognitive amplification occurs when a technological tool extends learners' intellectual performance while they retain responsibility for judgment, decision-making, and meaning construction (Salomon et al., 1991). In this form of engagement, students use ChatGPT to interrogate assumptions, compare perspectives, identify weaknesses, and refine their arguments. How these processes relate to ChatGPT use may vary with task design, learner agency, AI literacy, and the degree of dependency on the tool.

Existing studies provide valuable insights into students' acceptance of ChatGPT, their perceptions of its usefulness, and its contribution to writing support (Habibi et al., 2023; Strzelecki, 2024; Teng, 2024). Other studies have identified possible risks associated with frequent or poorly regulated AI use, including overreliance, weaker independent thinking, reduced creativity, and limited information

judgment (Abbas et al., 2024; Zhang et al., 2024). However, these perspectives have rarely been integrated within a single EFL classroom study. Accordingly, the present study examines both performance-based HOT outcomes and students' observable patterns of ChatGPT use within a structured instructional context. It compares structured ChatGPT-supported and conventional EFL instruction in terms of analysis, evaluation, and creation performance and examines students' ChatGPT-use patterns, with AI literacy and ChatGPT dependency included as complementary measures. Based on these objectives, the following research questions are addressed:

1. To what extent is structured ChatGPT-supported EFL learning associated with students' performance-based higher-order thinking in EFL argumentative writing, particularly in analysis, evaluation, and creation?
2. What reflective-evaluative, mixed, and product-oriented patterns of ChatGPT use were observed during the structured classroom tasks?

Literature Review

ChatGPT-Supported Argumentative Writing and Higher-Order Thinking

ChatGPT has become increasingly relevant to EFL learning because its dialogic interface enables learners to obtain immediate and individualized assistance. Unlike earlier automated writing tools that primarily identified linguistic errors, ChatGPT can respond to follow-up questions, elaborate on suggestions, and generate alternatives based on learners' prompts. EFL students have used the tool to develop ideas, clarify linguistic features, organize arguments, and revise academic texts (Meniado et al., 2024; Teng, 2024; Tram et al., 2024).

Empirical evidence suggests that the effectiveness of ChatGPT depends on its integration into purposeful learning activities. ChatGPT-supported feedback has been associated with improvements in academic and EFL writing (Mahapatra, 2024; Polakova & Ivenz, 2024). Its integration with collaboration scripts and argument mapping has also strengthened argumentative performance, critical-thinking awareness, and collaborative engagement (Darmawansah et al., 2025). Similarly, combining ChatGPT feedback with teacher guidance has improved multiple dimensions of EFL argumentative writing (Asadi et al., 2025).

Argumentative writing provides an appropriate context for assessing performance-based HOT because it requires learners to analyze issues, evaluate evidence and counterarguments, and construct coherent arguments (Li & Liu, 2021; Lu & Xie, 2019). Recent evidence indicates that GenAI is most likely to support critical thinking when its use is accompanied by structured guidance, critical evaluation, and sustained learner agency (Liu et al., 2024; Shen & Teng, 2024). Accordingly, the pedagogical value of ChatGPT should be assessed not only through performance gains but also through the extent to which learners retain intellectual responsibility during AI-mediated writing.

Cognitive Amplification, Cognitive Offloading, and Related Distinctions

Cognitive offloading and cognitive amplification provide complementary perspectives for examining how learners allocate intellectual effort when interacting with ChatGPT. Risko and Gilbert (2016) define cognitive offloading as the redistribution of cognitive demands to external resources. Although such redistribution may support task management, it becomes pedagogically problematic when AI-generated content substitutes for the analytical and evaluative processes that learners are expected to develop. Cognitive amplification, by contrast, occurs when technology extends learners' intellectual

capacity without diminishing their responsibility for judgment and meaning construction (Salomon et al., 1991). The distinction concerns not merely the frequency of ChatGPT use but the extent to which learners retain cognitive agency during AI-mediated tasks.

Empirical evidence suggests that the educational consequences of GenAI use are conditional rather than uniform. Abbas et al. (2024) reported that university students' use of GenAI may produce both beneficial and adverse outcomes, depending on the motivations and conditions underlying its use. Zhang et al. (2024) similarly associated problematic AI dependency with weaker independent thinking, reduced creativity, and lower information-judgment ability. In EFL writing, concerns have also emerged regarding plagiarism, overreliance, and the uncritical acceptance of generated content (Alkamel & Alwagieh, 2024; Mizumoto et al., 2024). The educational value of ChatGPT is shaped by how learners engage with AI-generated content and the extent to which they retain evaluative control. In this study, cognitive amplification and cognitive offloading provide theoretical lenses for interpreting students' observed patterns of ChatGPT use. Offloading can occur alongside evaluative engagement when students delegate some task demands while remaining responsible for the final response.

AI Literacy, Dependency, and Reflective Use

AI literacy offers an important lens for understanding differences in students' engagement with ChatGPT. Effective AI-mediated learning requires learners to understand AI systems, use them purposefully, evaluate their outputs, and make responsible ethical decisions (Ng et al., 2021). Wang et al. (2023) conceptualize AI literacy through four dimensions: awareness, use, evaluation, and ethics. Awareness concerns learners' understanding of AI functions and limitations; use refers to their ability to employ AI tools purposefully; evaluation involves assessing the accuracy, relevance, and quality of generated responses; and ethics encompasses responsible decisions related to transparency, academic integrity, and privacy.

These dimensions should be distinguished from ChatGPT dependency, which refers to excessive reliance on AI assistance for completing academic tasks and may weaken learner agency, independent task completion, and critical judgment (Goh et al., 2025; Zhang et al., 2024). Familiarity with ChatGPT does not necessarily indicate the ability to evaluate its output critically, just as frequent use does not automatically constitute meaningful cognitive engagement. In AI-assisted writing, reflective use involves examining AI-generated suggestions and deciding whether they should be accepted, modified, or rejected. Such evaluative decisions are central to maintaining learner agency during revision (Liu et al., 2024; Shen & Teng, 2024).

Pedagogical Scaffolding for Responsible and Evaluative AI Use

The educational value of ChatGPT also depends on how its use is structured within classroom tasks. In argumentative writing, scaffolding can require students to produce an initial response, examine alternative viewpoints, evaluate AI-generated feedback, and revise selectively rather than accept generated text directly. Research has shown that ChatGPT-supported writing is more productive when AI use is combined with teacher guidance, structured feedback, collaboration scripts, and critical evaluation (Asadi et al., 2025; Darmawansah et al., 2025; Liu et al., 2024). These practices position ChatGPT as a resource for reasoning and revision while keeping students responsible for their final decisions.

Research Gap and Rationale for the Present Study

Studies have examined students' acceptance and use of ChatGPT in higher education (Habibi et al., 2023; Shen & Teng, 2024; Strzelecki, 2024), its contribution to EFL and academic writing

(Asadi et al., 2025; Mahapatra, 2024; Meniado et al., 2024), and the risks of overreliance and uncritical AI use (Abbas et al., 2024; Liu et al., 2024; Mizumoto et al., 2024). However, limited evidence is available on whether structured ChatGPT-supported instruction is associated with performance-based HOT in supervised EFL argumentative writing completed without AI assistance. Moreover, prompt-response-revision patterns showing how students evaluate, modify, accept, or reject AI-generated suggestions remain underexamined. The present study addresses these gaps by examining HOT performance with students' observable patterns of ChatGPT use during a structured classroom intervention.

Methods

Research Design

This study employed a quasi-experimental mixed-methods design in which quantitative outcome data were complemented by ChatGPT interaction logs, classroom observations, and follow-up interviews (Creswell & Creswell, 2018). The quantitative phase used a quasi-experimental pretest-posttest control-group design with two intact classes. One class received structured ChatGPT-supported instruction, while the other completed comparable tasks through printed materials, peer discussion, lecturer guidance, and conventional feedback. Individual random assignment was not feasible because students were enrolled in fixed course sections. Both classes were taught by the same lecturer during the same term and received equivalent contact hours, learning objectives, topics, task loads, and assessment procedures.

The qualitative component consisted of three primary data sources with different functions: semi-structured interviews, ChatGPT interaction records, and classroom observation notes. Semi-structured interviews were conducted after the quantitative results and student profiles had been identified. During the instructional period, the experimental-group students submitted ChatGPT interaction records together with their revision records and brief reflection notes, while classroom observation notes were collected concurrently. These materials formed part of the ChatGPT interaction dataset. These qualitative materials were later analyzed to explain students' prompting behavior, revision decisions, classroom engagement, and reflective use of ChatGPT. Thus, the design followed explanatory sequential mixed-methods logic while incorporating embedded qualitative artifacts to support integration (Fetters et al., 2013).

Structured ChatGPT-Supported EFL Intervention

The study was conducted over 12 weeks, comprising 11 instructional weeks and a supervised posttest in Week 12. Throughout the intervention, students used ChatGPT powered by GPT-4o (OpenAI) through the ChatGPT web interface. The same model was used across the instructional activities to maintain consistency in the AI-supported learning condition. Both groups followed the same learning objectives, writing focus, workload, contact hours, and assessment schedule. The experimental group used ChatGPT with lecturer guidance, while the control group used printed materials, peer discussion, and lecturer feedback.

The intervention focused on higher-order EFL tasks requiring students to analyze argumentative texts, evaluate claims and evidence, and create revised arguments or problem-solution responses. Weeks 2–4 focused on argument analysis, Weeks 5–7 on evidence and claim evaluation, Weeks 8–10 on counterargument and rebuttal development, and Week 11 on integrated argumentative writing. Materials included lecturer-selected argumentative texts and university-related issues. Students produced analyses, evaluations, counterarguments, revised position paragraphs, and an argumentative essay.

In each session, students in the experimental group first produced an initial response independently before consulting ChatGPT. They used ChatGPT to request feedback, alternative viewpoints, counterarguments, or suggestions for improving the clarity and logic of their responses. After receiving ChatGPT output, students compared it with their initial work, selected relevant suggestions, rejected unsuitable ideas, and revised their final text in their own words. The lecturer modeled prompting, monitored AI use, and provided feedback on reasoning and revision. Students submitted their initial response, ChatGPT interaction record, revised response, and a brief reflection. After receiving a response, students could ask follow-up questions for clarification, compare the suggestions with their initial arguments, and then accept, modify, or reject them. For example, when ChatGPT identified weak evidence or an unclear claim, students revised the relevant section while retaining their original position.

The control group completed parallel tasks without ChatGPT, using printed materials, peer discussion, lecturer guidance, and conventional feedback. Students also produced initial and revised responses. AI-specific reflection notes were collected only from the experimental group as process evidence. The interaction logs captured how students applied the reflective prompting, comparison, and selective revision practices introduced during the intervention.

Setting and Participants

This study was conducted in the Information Systems Study Program at a private university in Lampung, Indonesia, where English is a compulsory undergraduate course. The setting was considered appropriate because students were familiar with digital technologies, making it relevant for investigating structured ChatGPT-supported EFL learning. Participants were 100 fourth-semester undergraduates enrolled during the 2025–2026 academic year. Purposive sampling was used to recruit students who had completed at least one previous English course, ensuring sufficient prior exposure to university-level EFL learning, and agreed to follow the study procedures. Participation was voluntary, and students were informed that withdrawal would not affect their grades.

The participants came from two intact classes and were divided into an experimental group (n = 50) and a control group (n = 50). The experimental group completed structured ChatGPT-supported tasks designed to foster analysis, evaluation, revision, and argument construction, whereas the control group completed comparable activities through printed materials, classroom discussion, lecturer guidance, and conventional feedback. The comparison involved two active instructional conditions rather than technology-supported learning versus no instruction. Both classes were allocated according to the existing timetable and taught by the same lecturer using parallel weekly plans, comparable materials, equivalent workloads, identical contact hours, and the same assessment schedule. Baseline comparability was examined using HOT pretest scores reported in Table 5 and demographic characteristics reported in Table 1.

For the qualitative follow-up phase, 20 students were selected from the experimental group using maximum variation sampling. The selection represented students with different HOT performance patterns and ChatGPT-use patterns, including reflective-evaluative dominant, mixed, and product-oriented dominant patterns. This follow-up sample enabled the study to examine how students planned, monitored, evaluated, and reflected on their use of ChatGPT during EFL argumentative tasks.

Instruments

For RQ1, performance-based HOT in EFL argumentative writing was assessed through supervised pretest and posttest essays completed without AI assistance. The prompts addressed related but

Table 1 *Participant demographic characteristics and prior ChatGPT use*

Characteristics	Experimental group (n = 50)	Control group (n = 50)	Total (n = 100)
Gender			
Male	27	26	53
Female	23	24	47
Age range			
18–19 years	9	10	19
20–21 years	34	33	67
22–23 years	7	7	14
Prior frequency of ChatGPT use			
Daily	8	7	15
Weekly	19	18	37
Monthly	12	13	25
Rarely	8	9	17
Never	3	3	6

non-identical issues: whether students should be permitted to use AI tools in academic writing and whether universities should regulate AI-generated content in assignments. Both tasks required students to analyze an issue, evaluate evidence, address an alternative viewpoint, and construct a justified position within 60 minutes and 300–400 words. Expert review and a pilot test with 30 students outside the main sample supported their comparability in clarity, cognitive demand, and completion time. Pilot participants produced comparable rubric scores across the two prompts, $t(29) = 0.84$, $p = .407$, and no significant difference in completion time was observed, $t(29) = 1.12$, $p = .272$.

Table 2 *Analytic rubric for HOT performance in EFL argumentative writing*

Dimension	5	4	3	2	1
Analysis	Examines the issue, relationships, assumptions, and tensions with nuance.	Identifies the issue and explains key relationships or assumptions.	Identifies the issue and provides basic connections among ideas.	Partially identifies the issue; connections among ideas remain limited.	Presents unclear information with no meaningful analysis.
Evaluation	Critically weighs evidence, alternatives, counterarguments, and limitations.	Evaluates evidence and addresses counterarguments with clear justification.	Uses relevant evidence and offers a basic judgment.	Provides limited evaluation and weak justification.	Offers unsupported judgments with no meaningful evaluation.
Creation	Synthesizes an original, coherent, and persuasive argument.	Develops a clear, organized, and well-supported argument.	Produces a generally coherent argument with sufficient support.	States a position but develops it inadequately.	Produces a fragmented response with no sustained position.

Essays were scored using an analytic rubric covering analysis, evaluation, and creation (Table 2). Three experts evaluated the prompts and rubric descriptors, yielding Aiken's V values of .89 and .91, respectively. Two trained raters independently scored all essays, and final scores were calculated by averaging their ratings. Inter-rater reliability was estimated using ICC(2,k): .86 for analysis, 95% CI [.80, .90]; .88 for evaluation, 95% CI [.83, .92]; .87 for creation, 95% CI [.82, .91]; and .90 for the total score, 95% CI [.86, .93]. Because the rubric was applied to argumentative essays, the scores represent performance-based HOT in EFL argumentative writing rather than general HOT ability.

Both the experimental and control groups completed the 11-item ChatGPT Dependency Scale adapted from Goh et al. (2025) and the 12-item AI Literacy Scale adapted from Wang et al. (2023) at baseline and after the intervention. The instruments assessed dependency, AI awareness, use, evaluation, and ethics using five-point Likert scales. Items were contextualized for EFL learning, translated into Indonesian, independently back-translated, and reviewed by three experts. Cronbach's alpha coefficients were .88 for dependency and .91 for AI literacy at baseline and .90 and .92, respectively, after the intervention. Subscale coefficients ranged from .81 to .89. Subscale and total scores were calculated as item means. The questionnaire scores provided complementary descriptive information on students' AI literacy and ChatGPT dependency across the study period.

Data Collection

Data were collected in three stages. Before the intervention, all participants completed a supervised HOT argumentative essay pretest, the AI Literacy Scale, and the ChatGPT Dependency Scale. During the 11 instructional weeks, the experimental group completed structured ChatGPT-supported tasks, while the control group completed comparable activities through printed materials, peer discussion, lecturer guidance, and conventional feedback. Experimental-group students also submitted interaction logs, revision records, and brief reflections explaining how they evaluated and used ChatGPT suggestions. In Week 12, all participants completed a comparable supervised HOT argumentative essay posttest and completed the AI Literacy Scale and ChatGPT Dependency Scale again. ChatGPT and other AI tools were prohibited during both the pretest and posttest. Baseline and post-intervention questionnaire scores were used to characterize students' AI-related dispositions and examine descriptive changes across the intervention period. Questionnaire scores were treated as secondary descriptive evidence.

The study comprised 11 instructional weeks followed by a supervised posttest in Week 12. Week 1 introduced ChatGPT ethics and prompting; Weeks 2–4 focused on analysis; Weeks 5–7 on evaluation; Weeks 8–10 on counterargument and creation; and Week 11 on integrated essay writing. In Week 12, all participants completed the supervised posttest without AI assistance. The detailed study schedule is presented in Table 3.

Data Analysis

Data analysis proceeded in two stages. First, essay scores were screened for missing values, outliers, normality, and homogeneity of variance. Baseline equivalence was examined using independent-samples t-tests. To answer RQ1, ANCOVA compared posttest HOT performance while controlling for pretest scores. Separate models were estimated for the total score and the dimensions of analysis, evaluation, and creation. Holm-adjusted p-values were reported for dimension-level comparisons, partial eta squared was used to describe the magnitude of the adjusted group differences, and HOT gain was calculated by subtracting pretest from posttest scores. Because each condition comprised one intact class, instructional condition and class membership could not be separated. The ANCOVA therefore estimated adjusted between-class differences at the individual level, and these differences were interpreted associatively rather than causally.

Table 3 *Structure of the 12-week study*

Week	Focus and learning objective	Materials/written product	Experimental group	Control group	Teacher support/ feedback	Reflection and time
1	Orientation to argumentative writing, AI ethics, and prompting	Practice argumentative response	ChatGPT ethics and prompt modelling	Lecturer explanation without AI	Explained tasks, criteria, and procedures	Orientation; equivalent class time
2	Argument analysis	Argumentative text; initial and revised analysis	Initial response -> ChatGPT feedback -> revision	Initial response -> peer discussion -> revision	Guided identification of issues and assumptions	Experimental: brief reflection; equivalent class time
3	Argument analysis	Argumentative text; initial and revised analysis	ChatGPT-assisted examination of argument weaknesses	Printed text and peer discussion	Feedback on reasoning and relationships among ideas	Experimental: brief reflection; equivalent class time
4	Argument analysis	Argumentative text; revised argument paragraph	ChatGPT feedback on weaknesses -> revision	Lecturer/peer feedback -> revision	Feedback on argument weaknesses and revision	Experimental: brief reflection; equivalent class time
5	Evidence and claim evaluation	Argumentative text; evaluation response	AI-assisted comparison -> selective revision	Lecturer-guided evaluation -> revision	Guided evaluation of claims and evidence	Experimental: brief reflection; equivalent class time
6	Alternative viewpoints	University-related issue; revised response	ChatGPT-generated alternatives -> evaluation -> revision	Peer discussion of alternatives -> revision	Guided comparison and justification	Experimental: brief reflection; equivalent class time
7	Argument strength	Argumentative text; evaluative paragraph	ChatGPT-assisted evaluation -> revision	Lecturer-guided evaluation -> revision	Feedback on evidence and justification	Experimental: brief reflection; equivalent class time
8	Counterargument development	University-related issue; counterargument paragraph	ChatGPT-supported counterargument -> evaluation -> revision	Peer discussion -> counterargument revision	Guided development of relevant counterarguments	Experimental: brief reflection; equivalent class time
9	Counterargument and rebuttal	Argumentative task; counterargument and rebuttal	ChatGPT-supported exploration -> revision	Peer/lecturer feedback -> revision	Feedback on relevance and logic	Experimental: brief reflection; equivalent class time
10	Revised position development	Argumentative issue; revised position paragraph	ChatGPT-supported revision	Conventional feedback and revision	Feedback on coherence and justification	Experimental: brief reflection; equivalent class time
11	Integrated argumentative writing	Full argumentative essay; initial and revised drafts	ChatGPT-supported feedback and revision	Lecturer feedback and revision	Feedback on argument development and final revision	Experimental: brief reflection; equivalent class time
12	Posttest	300–400-word supervised argumentative essay	No AI	No AI	Same assessment procedures	No reflection; 60 minutes

To answer RQ2, ChatGPT interaction logs, revision records, reflection notes, interview transcripts, and observation notes were analyzed using a hybrid deductive-inductive procedure. During each of the ten core sessions from Weeks 2 to 11, experimental-group students were required to export and submit the complete ChatGPT interaction record associated with the assigned focal task. The submitted record included the initial prompt, all follow-up prompts, ChatGPT responses, and the corresponding revision evidence. Students were not permitted to select only particular exchanges from a conversation thread. This procedure reduced the risk of selective submission and enabled the analysis to capture both reflective-evaluative and product-oriented patterns of ChatGPT use.

For coding purposes, the unit of analysis was a traceable prompt-response-revision sequence linked to the assigned focal task. Each task-specific interaction record was treated as one coded sequence. When a record contained multiple turns, the initial prompt, follow-up prompts, ChatGPT responses, and linked revision were read together and coded as a single sequence because they addressed the same focal task. Follow-up turns informed the coding of the sequence rather than being counted as separate units. Of the 500 required task-specific interaction records, 486 were submitted (97.2%). Thirty-eight students submitted all ten records, ten submitted nine records, and two submitted eight records because of absence during one or two sessions. Fourteen records were excluded because the ChatGPT response could not be linked reliably to a subsequent revision decision. Consequently, 472 valid prompt-response-revision sequences were retained, representing 94.4% of the required records and 97.1% of the submitted records. All valid records were analyzed; no student-selected sampling of interaction exchanges was applied.

Each of the 472 valid prompt-response-revision sequences was independently coded by two researchers using a predefined codebook. Before coding the main dataset, the coders reviewed the operational definitions and independently coded 50 practice sequences drawn randomly from the valid records. Disagreements in the calibration subset were discussed to refine the codebook and clarify borderline cases. Both researchers then independently coded all 472 sequences before any consensus discussion, yielding an initial agreement of 91.3% and a Cohen's κ of .81, indicating almost perfect agreement. Disagreements were resolved through discussion and re-examination of the original prompt, ChatGPT response, linked revision record, and reflection note.

A sequence was coded as reflective-evaluative when the student used ChatGPT to examine assumptions, compare viewpoints, evaluate evidence, identify weaknesses, generate counterarguments, or improve reasoning, and the revision trace showed selective acceptance, modification, or rejection of AI-generated suggestions based on their relevance, accuracy, or argumentative value. A sequence was coded as product-oriented when the student requested a complete answer, ready-to-submit paragraph, or direct task completion and reproduced the AI-generated response with limited evaluation. Direct copying referred to reproducing AI-generated text with little or no change. Minimal transformation included lexical, grammatical, or formatting changes without substantive restructuring, whereas substantive revision involved changes to reasoning, evidence, organization, or argumentative position. Borderline cases involving partial copying or limited modification were coded as reflective-evaluative only when a clear evaluative decision or substantive transformation was traceable; otherwise, they were coded as product-oriented. Reflection notes, interview transcripts, and observation notes were used to contextualize the coded patterns.

ChatGPT-use patterns were classified based on the proportion of coded prompt-response-revision sequences. The classification was based on observable interaction and revision behaviours, while cognitive amplification (Salomon et al., 1991) and cognitive offloading (Risko & Gilbert, 2016) informed the interpretation of these patterns. A provisional 60% threshold was set a priori as the analytical criterion for identifying a dominant pattern. Students with at least 60% reflective-evaluative sequences

were classified as reflective-evaluative dominant, those with at least 60% product-oriented sequences as product-oriented dominant, and the remainder as mixed.

AI literacy and ChatGPT dependency scores were summarized descriptively and analyzed separately from the classification of ChatGPT-use patterns. Quantitative and qualitative findings were integrated through a joint display. The operational criteria are presented in Table 4.

Table 4 *Operational criteria for ChatGPT-use patterns*

Behavioral pattern	Criterion	Observable pattern
Reflective-evaluative dominant	≥60% reflective-evaluative sequences	Evaluated AI output, compared viewpoints, identified weaknesses, and revised selectively.
Mixed	Neither dominant criterion reached	Shifted between reflective evaluation and direct task-completion use.
Product-oriented dominant	≥60% product-oriented sequences	Requested complete responses or ready-to-submit text and made limited substantive revision.

Note: The categories summarize observable patterns of ChatGPT use during the classroom tasks. The 60% threshold was used as a provisional a priori criterion for identifying a dominant pattern and requires further validation. Cognitive amplification and cognitive offloading were used to interpret these patterns theoretically.

Ethical Considerations

Ethical approval was obtained from the Ethics Committee of Institut Bina Nusantara under ethical clearance number 024/IBN/LPPM/U/I/2026. All participants provided written informed consent before data collection and were informed that participation was voluntary, withdrawal was permitted at any stage, and their decisions would not affect their academic grades. Participant identities were replaced with codes in all datasets, transcripts, quotations, and reports. ChatGPT interaction logs and revision records were de-identified and stored in password-protected files accessible only to the research team. Students were instructed not to enter names, student numbers, contact details, or other personal information into ChatGPT during the intervention. Anonymized data may be made available upon reasonable request, subject to institutional ethical requirements and participant confidentiality. The authors declare that they have no conflicts of interest.

Results

Structured ChatGPT-Supported EFL Learning and Higher-Order Thinking Performance (RQ1)

The first research question examined whether students receiving structured ChatGPT-supported EFL instruction achieved higher adjusted HOT performance in argumentative writing than students receiving conventional EFL instruction. Performance was assessed using an analytic rubric consisting of analysis, evaluation, and creation. Each dimension was scored from 1 to 5, producing a total score ranging from 3 to 15.

As shown in Table 5, both groups improved, but the experimental group showed larger gains across all HOT dimensions, with the largest gain in creation. The groups were comparable at baseline across total HOT performance and the three HOT dimensions.

Table 5 *Descriptive statistics of performance-based HOT pretest, posttest, and gain scores*

Variable	Group	Pretest M	Pretest SD	Posttest M	Posttest SD	Gain
Total HOT performance	Experimental	8.72	1.08	11.46	1.14	2.74
	Control	8.64	1.11	9.76	1.18	1.12
Analysis	Experimental	2.96	0.42	3.78	0.47	0.82
	Control	2.92	0.44	3.31	0.48	0.39
Evaluation	Experimental	2.83	0.46	3.72	0.49	0.89
	Control	2.81	0.45	3.18	0.46	0.37
Creation	Experimental	2.93	0.43	3.96	0.48	1.03
	Control	2.91	0.44	3.27	0.47	0.36

Table 6 *ANCOVA results for performance-based HOT posttest scores controlling for pretest scores*

Dependent variable	Group	Estimated marginal M	SE	95% CI	F (1, 97)	p	Holm-adjusted p	ηp^2
Total HOT performance	Experimental	11.43	0.16	[11.11, 11.75]	55.28	< .001	—	.363
	Control	9.79	0.16	[9.47, 10.11]				
Analysis	Experimental	3.77	0.07	[3.63, 3.91]	29.64	< .001	< .001	.234
	Control	3.32	0.07	[3.18, 3.46]				
Evaluation	Experimental	3.71	0.07	[3.57, 3.85]	35.17	< .001	< .001	.266
	Control	3.19	0.07	[3.05, 3.33]				
Creation	Experimental	3.94	0.07	[3.80, 4.08]	48.63	< .001	< .001	.334
	Control	3.29	0.07	[3.15, 3.43]				

Note: Holm adjustment was applied to the three dimension-level comparisons. The total HOT score was treated as the primary outcome.

As shown in Table 6, ANCOVA showed a significant adjusted difference in total HOT performance between the two classes after controlling for pretest scores, $F(1, 97) = 55.28$, $p < .001$, $\eta p^2 = .363$. The experimental group achieved a higher adjusted posttest score than the control group. Significant differences were also found for analysis, evaluation, and creation after Holm adjustment. Among the three HOT dimensions, the largest adjusted difference was observed for creation ($\eta p^2 = .334$), followed by evaluation ($\eta p^2 = .266$) and analysis ($\eta p^2 = .234$).

Patterns of ChatGPT Use During the Structured Intervention (RQ2)

The second research question examined the patterns of ChatGPT use observed during the structured classroom tasks. Table 7 integrates the distribution of the three behavioral patterns with the corresponding prompt-response-revision sequences, revision practices, and representative student accounts.

Reflective-evaluative use was the most common pattern, with 31 students (62%) classified as reflective-evaluative dominant, followed by 12 students (24%) showing mixed use and seven students (14%) showing product-oriented dominant use. Across the 472 valid sequences, reflective-evaluative engagement was more common than product-oriented engagement. Students showing reflective-evaluative

Table 7 *Joint display of ChatGPT-use patterns and revision practices*

ChatGPT-use pattern	Students, n (%)	Valid sequences, n	Reflective-evaluative sequences, n (%)	Product-oriented sequences, n (%)	Revision practice	Representative quotation
Reflective-evaluative dominant	31 (62%)	293	236 (80.5%)	57 (19.5%)	Students evaluated AI suggestions, compared viewpoints, rejected unsuitable responses, and revised selectively.	"I asked ChatGPT to show the weakness of my paragraph, but I did not copy it directly because some examples were not related to my topic" (S07).
Mixed	12 (24%)	113	64 (56.6%)	49 (43.4%)	Students shifted between reflective evaluation and direct task-completion use across tasks.	"Sometimes I used ChatGPT to get ideas, but when the task was difficult, I asked it to make the paragraph for me first" (S14).
Product-oriented dominant	7 (14%)	66	14 (21.2%)	52 (78.8%)	Students more frequently requested complete responses and made limited substantive revisions.	"I used ChatGPT because it was faster. I usually followed the answer because I was not sure how to change it" (S20).
Total	50 (100%)	472	314 (66.5%)	158 (33.5%)	—	—

Note: Percentages for reflective-evaluative and product-oriented sequences were calculated within each behavioral pattern. Behavioral-pattern classification was conducted at the student level using the criteria described in the Methods.

patterns typically examined AI-generated suggestions, compared alternatives, and revised selectively, whereas product-oriented use more frequently involved requests for complete responses and limited substantive revision. Mixed-pattern students shifted between these forms of engagement across tasks. These distributions describe observable patterns of ChatGPT use during the structured classroom activities.

AI literacy and ChatGPT dependency were included as secondary descriptive measures alongside the behavioral-pattern analysis. Table 8 presents pretest and posttest questionnaire scores for the experimental and control groups.

As shown in Table 8, the experimental group showed larger descriptive increases across all AI literacy dimensions, particularly AI evaluation, while ChatGPT dependency decreased slightly. The control group showed comparatively modest changes across the same measures.

Table 8 *Secondary descriptive questionnaire scores by group and time*

Measure	Group	Pretest M	Pretest SD	Posttest M	Posttest SD	Mean Change
AI awareness	Experimental	3.46	0.55	3.82	0.49	+0.36
	Control	3.43	0.54	3.50	0.52	+0.07
AI use	Experimental	3.31	0.61	3.86	0.54	+0.55
	Control	3.29	0.59	3.38	0.57	+0.09
AI evaluation	Experimental	3.20	0.58	3.92	0.50	+0.72
	Control	3.18	0.57	3.32	0.55	+0.14
AI ethics	Experimental	3.55	0.52	3.88	0.47	+0.33
	Control	3.51	0.53	3.57	0.50	+0.06
Overall AI literacy	Experimental	3.38	0.30	3.87	0.22	+0.49
	Control	3.35	0.31	3.44	0.29	+0.09
ChatGPT dependency	Experimental	2.57	0.56	2.43	0.51	–0.14
	Control	2.55	0.55	2.52	0.53	–0.03

Note: All questionnaire scores were calculated on five-point scales. Mean change was calculated as posttest minus pretest. Both groups completed the AI Literacy Scale and ChatGPT Dependency Scale before and after the intervention. Questionnaire results were treated as secondary descriptive measures.

Discussion

This study examined whether structured ChatGPT-supported EFL learning was associated with students' performance-based higher-order thinking in argumentative writing and how students engaged with ChatGPT during the structured intervention. The findings are interpreted in relation to the structured instructional package in which ChatGPT was combined with guided prompting, evaluation, revision, and lecturer support.

The first finding showed that students in the experimental group achieved significantly higher HOT performance than those in the control group after pretest scores were controlled. The difference was found across analysis, evaluation, and creation, with creation showing the largest adjusted difference among the three HOT dimensions. The experimental condition combined ChatGPT access with independent initial responses, lecturer modelling, guided comparison, selective revision, and reflection. Students were required to identify weaknesses in their arguments, compare alternative viewpoints, examine possible rebuttals, and revise their final texts. The higher HOT scores occurred within this combination of structured activities, lecturer guidance, and feedback. Similar benefits of structured ChatGPT integration have been reported for writing and argumentation when AI support is combined with instructional guidance rather than used independently (Asadi et al., 2025; Darmawansah et al., 2025; Mahapatra, 2024).

The largest adjusted difference occurred for creation, showing that the between-class difference extended beyond analysis and evaluation to the construction of original responses. This pattern is consistent with Asadi et al. (2025), who found that ChatGPT feedback combined with instructional support improved EFL argumentative writing. Liu et al. (2024) similarly reported that GenAI-assisted composing involved evaluative decisions and the development of new textual elements. The higher creation scores were observed alongside AI-supported exploration, repeated revision, and instructional scaffolding.

The log data showed that reflective-evaluative use was more common than product-oriented use during the structured intervention. Students frequently examined assumptions, compared viewpoints, requested counterarguments, and selectively revised AI-generated suggestions. These behaviours emerged within supervised, teacher-designed tasks that explicitly emphasized prompting, evaluation, and revision, and therefore reflect students' engagement within this instructional context. Previous research likewise shows that structured and guided ChatGPT use can support idea development, evaluation, and revision in EFL writing (Meniado et al., 2024; Shen & Teng, 2024; Werdiningsih et al., 2024). Reflective-evaluative use was broadly consistent with cognitive amplification because students retained control over judgment and revision, whereas product-oriented use was more consistent with cognitive offloading because AI played a larger role in task completion (Risko & Gilbert, 2016; Salomon et al., 1991). These interpretations are based on the observed interaction and revision patterns.

A smaller group of students showed product-oriented dominant use. These students used ChatGPT mainly to save time and complete tasks. Their prompts frequently requested complete responses, while their final texts showed limited revision. Related concerns have been reported by Abbas et al. (2024), who linked frequent ChatGPT use with procrastination, memory loss, and lower academic performance. Zhang et al. (2024) also found that problematic AI use was associated with weaker independent thinking and lower information-judgment ability. In the present study, these behaviours reflect greater reliance on AI-generated products during the structured classroom tasks and are consistent with greater cognitive offloading.

The mixed-use pattern provides another important insight. Some students moved between reflective and more direct task-completion use across tasks. This variation suggests that ChatGPT-use behaviour was responsive to task and classroom conditions rather than representing a fixed learner characteristic. Task difficulty, time constraints, confidence, and instructional requirements may have influenced how students engaged with the tool. Zaim et al. (2024) similarly emphasized that GenAI use is shaped by instructional purposes, rules, and learning activities. The process analysis showed how students engaged with ChatGPT during the supported classroom activities.

The study contributes to the literature by combining performance-based assessment of HOT with interaction and revision evidence of students' ChatGPT use. HOT was assessed through argumentative essays scored for analysis, evaluation, and creation rather than through student perceptions alone. ChatGPT interaction records and linked revision evidence were used to identify reflective-evaluative, mixed, and product-oriented patterns, which were interpreted in relation to cognitive amplification and cognitive offloading. Reflection notes, interviews, and classroom observations further contextualized these patterns. Therefore, the findings show that students engaged with ChatGPT in different ways within the same structured classroom setting and that their engagement varied across task demands and revision practices.

Beyond these theoretical considerations, the findings offer practical implications for EFL instruction. EFL lecturers can structure ChatGPT-supported activities by requiring students to develop an initial response before using ChatGPT, document their prompts, evaluate AI-generated suggestions, explain revision decisions, and submit evidence of changes made to their texts. Prompt training can emphasize counterarguments, evidence checking, alternative viewpoints, and weak assumptions rather than requests for complete answers. Reflection notes and revision records can help lecturers examine how students interact with AI. The present findings suggest that structured prompting, evaluation, and revision can provide a useful framework for integrating ChatGPT while maintaining students' responsibility for final decisions.

Conclusion

This study showed that structured ChatGPT-supported EFL learning was associated with higher performance-based HOT in argumentative writing in the participating classes. After pretest scores were

controlled, the experimental group achieved higher posttest scores than the control group across analysis, evaluation, and creation, with creation showing the largest dimension-level difference. The experimental condition combined ChatGPT-supported activities with guided prompting, evaluation, revision, reflection, and lecturer support. Most students in the experimental group showed reflective-evaluative use during the structured tasks, whereas a smaller group showed predominantly product-oriented use. These patterns illustrate different forms of engagement with ChatGPT within the classroom context and are best understood as context-dependent behaviours rather than fixed learner characteristics.

Because each instructional condition involved one intact class, the group differences are interpreted within the context of the participating classes. Process evidence was collected only from the experimental group, and the interaction logs were generated through teacher-designed tasks that explicitly emphasized evaluation and revision. These conditions should be considered when interpreting the observed ChatGPT-use patterns and questionnaire responses. The 60% threshold was used as a provisional a priori analytical criterion and requires further validation across samples and tasks. Future studies could extend this work across multiple classes and institutions, include comparable process evidence across conditions, and examine whether the identified behavioural patterns remain consistent across tasks and over time. Additional process measures could also examine more closely how these behavioural patterns correspond to cognitive amplification and cognitive offloading.

Acknowledgments

The authors thank the Directorate General of Higher Education, Ministry of Education and Culture of the Republic of Indonesia, for providing financial support for this research under Contract No. 94/DST/C/AL.04.02/2026, dated April 9, 2026.

Declaration of AI and AI-Assisted Technologies

While preparing this work, the authors used GenAI tools and Grammarly to support language editing and readability. The authors critically reviewed and revised all AI-assisted outputs and take full responsibility for the final content.

References

Abbas, M., Jam, F. A., & Khan, T. I. (2024). Is it harmful or helpful? Examining the causes and consequences of generative AI usage among university students. *International Journal of Educational Technology in Higher Education*, *21*, Article 10. https://doi.org/10.1186/s41239-024-00444-7

Alkamel, M. A. A., & Alwagieh, N. A. S. (2024). Utilizing an adaptable artificial intelligence writing tool (ChatGPT) to enhance academic writing skills among Yemeni university EFL students. *Social Sciences & Humanities Open*, *10*, 101095. https://doi.org/10.1016/j.ssaho.2024.101095

Andewi, W., Waziana, W., Wibisono, D., Putra, K. A., Hastomo, T., & Oktarin, I. B. (2025). From prompting to proficiency: A mixed-methods analysis of prompting with ChatGPT versus lecturer interaction in an EFL classroom. *Journal of Studies in the English Language*, *20*(2), 210–238. https://doi.org/10.64731/jsel.v20i2.282318

Asadi, M., Ebadi, S., & Mohammadi, L. (2025). The impact of integrating ChatGPT with teachers' feedback on EFL writing skills. *Thinking Skills and Creativity*, *56*, 101766. https://doi.org/10.1016/j.tsc.2025.101766

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage .

Darmawansah, D., Rachman, D., Febiyani, F., & Hwang, G.-J. (2025). ChatGPT-supported collaborative argumentation: Integrating collaboration script and argument mapping to enhance

EFL students' argumentation skills. *Education and Information Technologies*, *30*(3), 3803–3827. https://doi.org/10.1007/s10639-024-12986-4

Fetters, M. D., Curry, L. A., & Creswell, J. W. (2013). Achieving integration in mixed methods designs—Principles and practices. *Health Services Research*, *48*, 2134–2156. https://doi.org/10.1111/1475-6773.12117

Goh, A. Y. H., Hartanto, A., & Majeed, N. M. (2025). Generative artificial intelligence dependency: Scale development, validation, and its motivational, behavioral, and psychological correlates. *Computers in Human Behavior Reports*, *20*, 100845. https://doi.org/10.1016/j.chbr.2025.100845

Habibi, A., Muhaimin, M., Danibao, B. K., Wibowo, Y. G., Wahyuni, S., & Octavia, A. (2023). ChatGPT in higher education learning: Acceptance and use. *Computers and Education: Artificial Intelligence*, *5*, 100190. https://doi.org/10.1016/j.caeai.2023.100190

Hastomo, T., Hamidah, F. N., Nurchurifiani, E., S, K., Hakim, L. N., Nasution, S. S., & Muliyah, P. (2026). EFL teachers as intercultural gatekeepers: Designing culturally responsive materials with GenAI in Indonesia. *Language Education & Assessment*, *9*, 104539. https://doi.org/10.29140/lea.2026.104539

Li, X., & Liu, J. (2021). Mapping the taxonomy of critical thinking ability in EFL. *Thinking Skills and Creativity*, *41*, 100880. https://doi.org/10.1016/j.tsc.2021.100880

Liang, W., & Wu, Y. (2024). Exploring the use of ChatGPT to foster EFL learners' critical thinking skills from a post-humanist perspective. *Thinking Skills and Creativity*, *54*, 101645. https://doi.org/10.1016/j.tsc.2024.101645

Liu, M., Zhang, L. J., & Biebricher, C. (2024). Investigating students' cognitive processes in generative AI-assisted digital multimodal composing and traditional writing. *Computers & Education*, *211*, 104977. https://doi.org/10.1016/j.compedu.2023.104977

Lu, D., & Xie, Y. (2019). The effects of a critical thinking oriented instructional pattern in a tertiary EFL argumentative writing course. *Higher Education Research & Development*, *38*(5), 969–984. https://doi.org/10.1080/07294360.2019.1607830

Mahapatra, S. (2024). Impact of ChatGPT on ESL students' academic writing skills: A mixed methods intervention study. *Smart Learning Environments*, *11*, Article 9. https://doi.org/10.1186/s40561-024-00295-9

Meniado, J. C., Huyen, D. T. T., Panyadilokpong, N., & Lertkomolwit, P. (2024). Using ChatGPT for second language writing: Experiences and perceptions of EFL learners in Thailand and Vietnam. *Computers and Education: Artificial Intelligence*, *7*, 100313. https://doi.org/10.1016/j.caeai.2024.100313

Mizumoto, A., Yasuda, S., & Tamura, Y. (2024). Identifying ChatGPT-generated texts in EFL students' writing: Through comparative analysis of linguistic fingerprints. *Applied Corpus Linguistics*, *4*(3), 100106. https://doi.org/10.1016/j.acorp.2024.100106

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence*, *2*, 100041. https://doi.org/10.1016/j.caeai.2021.100041

Polakova, P., & Ivenz, P. (2024). The impact of ChatGPT feedback on the development of EFL students' writing skills. *Cogent Education*, *11*(1), Article 2410101. https://doi.org/10.1080/2331186X.2024.2410101

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, *20*(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002

Salomon, G., Perkins, D., & Globerson, T. (1991). Partners in cognition: Extending human intelligence with intelligent technologies. *Educational Researcher*, *20*(3), 2–9. https://doi.org/10.3102/0013189X020003002

Shen, X., & Teng, M. F. (2024). Three-wave cross-lagged model on the correlations between critical thinking skills, self-directed learning competency and AI-assisted writing. *Thinking Skills and Creativity*, *52*, 101524. https://doi.org/10.1016/j.tsc.2024.101524

Strzelecki, A. (2024). Students' acceptance of ChatGPT in higher education: An extended unified theory of acceptance and use of technology. *Innovative Higher Education*, *49*(2), 223–245. https://doi.org/10.1007/s10755-023-09686-1

Teng, M. F. (2024). "ChatGPT is the companion, not enemies": EFL learners' perceptions and experiences in using ChatGPT for feedback in writing. *Computers and Education: Artificial Intelligence*, *7*, 100270. https://doi.org/10.1016/j.caeai.2024.100270

Tram, N. H. M., Nguyen, T. T., & Tran, C. D. (2024). ChatGPT as a tool for self-learning English among EFL learners: A multi-methods study. *System*, *127*, 103528. https://doi.org/10.1016/j.system.2024.103528

Wang, B., Rau, P.-L. P., & Yuan, T. (2023). Measuring user competence in using artificial intelligence: Validity and reliability of artificial intelligence literacy scale. *Behaviour & Information Technology*, *42*(9), 1324–1337. https://doi.org/10.1080/0144929X.2022.2072768

Werdiningsih, I., Marzuki, & Rusdin, D. (2024). Balancing AI and authenticity: EFL students' experiences with ChatGPT in academic writing. *Cogent Arts & Humanities*, *11*(1), Article 2392388. https://doi.org/10.1080/23311983.2024.2392388

Zaim, M., Arsyad, S., Waluyo, B., Ardi, H., Al Hafizh, M., Zakiyah, M., Syafitri, W., Nusi, A., & Hardiah, M. (2024). AI-powered EFL pedagogy: Integrating generative AI into university teaching preparation through UTAUT and activity theory. *Computers and Education: Artificial Intelligence*, *7*, 100335. https://doi.org/10.1016/j.caeai.2024.100335

Zhang, S., Zhao, X., Zhou, T., & Kim, J. H. (2024). Do you have AI dependency? The roles of academic self-efficacy, academic stress, and performance expectations on problematic AI usage behavior. *International Journal of Educational Technology in Higher Education*, *21*(1), Article 34. https://doi.org/10.1186/s41239-024-00467-0

Appendix A

Examples of ChatGPT Interaction-Log Coding

The unit of analysis was one task-specific prompt-response-revision sequence. When an interaction contained multiple turns, the initial prompt, follow-up prompts, ChatGPT responses, and linked revision were treated as one sequence when they addressed the same focal task. Follow-up turns therefore contributed to the interpretation of the sequence rather than being counted as separate units.

Table A1 *Reflective-evaluative sequence*

Element	Record
Student	S07
Week	6
Task focus	Evaluating alternative viewpoints
Initial response	Universities should regulate students' use of AI in academic writing because students may become too dependent on technology. However, banning AI completely is not appropriate because students can use it to improve their ideas and language. Universities should provide clear rules so students know which uses are acceptable.
Student prompt	I wrote this paragraph about whether universities should regulate AI use in academic writing. Please identify the weaknesses in my argument. Do not rewrite the paragraph. Focus on the logic, evidence, and alternative viewpoints.
ChatGPT response	Your position is clear, but the argument could be strengthened in three areas: 1. The claim that students may become dependent on AI is stated without explanation or evidence. 2. The paragraph mentions that AI can improve ideas and language, but it does not explain how this benefit can coexist with regulation. 3. An opposing viewpoint is missing. For example, some students may argue that regulation limits access to useful learning tools. You could also refer to how AI is used in professional workplaces, although that example may be less relevant to a paragraph focused on university academic writing. You could strengthen the paragraph by explaining what type of regulation would allow educational use while preventing inappropriate dependence.
Follow-up prompt	The example about professional workplaces is not related to my topic. What counterargument would fit my position better?
ChatGPT follow-up response	A more suitable counterargument would be that regulation may discourage students from experimenting with useful AI tools. You could acknowledge this concern and then argue that clear guidelines are different from strict restrictions. Guidelines can allow students to use AI for brainstorming or feedback while requiring them to produce and justify their own final work.
Linked revision	Universities should regulate students' use of AI in academic writing because excessive reliance may reduce students' responsibility for developing and evaluating their own arguments. However, regulation should not prevent students from using AI for brainstorming, language support, or feedback. Although some students may view regulation as limiting access to useful technology, clear guidelines can distinguish acceptable learning support from inappropriate substitution of students' own work. Universities should therefore establish transparent rules that allow educational use while maintaining students' responsibility for the final submission.
Coding rationale	The sequence was coded as reflective-evaluative because the student asked ChatGPT to identify weaknesses rather than produce a replacement paragraph, rejected an example that was not relevant to the focal topic, requested a more suitable counterargument, and selectively incorporated the resulting feedback. The linked revision substantially developed the reasoning and counterargument while retaining the student's original position.
Supporting reflection	"I asked ChatGPT to show the weakness of my paragraph, but I did not copy it directly because some examples were not related to my topic" (S07).

Table A2 *Product-oriented sequence*

Element	Record
Student	S20
Week	10
Task focus	Developing a revised position paragraph
Initial response	AI can help university students, but sometimes students use it too much. Universities need rules about AI.
Student prompt	Write a complete academic paragraph about why universities should regulate AI-generated content in student assignments. Make it around 130 words and include a clear argument and counterargument.
ChatGPT response	Universities should regulate AI-generated content in student assignments to protect academic integrity and ensure that students remain responsible for their own learning. Although generative AI can support brainstorming, language improvement, and idea development, excessive reliance on AI-generated text may make it difficult for lecturers to determine whether submitted work reflects students' actual abilities. Some students may argue that regulation limits access to useful educational technology; however, regulation does not necessarily mean prohibition. Universities can establish clear guidelines that permit AI for appropriate learning support while requiring students to disclose its use and independently evaluate generated content. Such policies can preserve the educational benefits of AI while reducing inappropriate dependence and maintaining meaningful assessment of students' knowledge and skills.
Follow-up prompt	Make it simpler and easier for an EFL student.
ChatGPT follow-up response	Universities should regulate AI-generated content in student assignments because students need to remain responsible for their own learning. AI can help students find ideas, improve language, and receive feedback, but using too much AI-generated text can make it difficult for teachers to know students' real abilities. Some students may think that AI rules will limit useful technology. However, regulation does not mean that AI must be banned. Universities can allow students to use AI for learning support while asking them to explain how they used it and check the information carefully. Clear rules can help students receive the benefits of AI without depending on it too much and can also protect academic integrity.
Linked revision	Universities should regulate AI-generated content in student assignments because students need to remain responsible for their own learning. AI can help students find ideas, improve language, and receive feedback, but using too much AI-generated text can make it difficult for teachers to know students' real abilities. Some students may think that AI rules will limit useful technology. However, regulation does not mean that AI must be banned. Universities can allow students to use AI for learning support while asking them to explain how they used it and check the information carefully. Clear rules can help students receive the benefits of AI without depending on it too much and can protect academic integrity.
Coding rationale	The sequence was coded as product-oriented because the student requested a complete paragraph and subsequently asked ChatGPT to simplify the generated text. The submitted revision closely followed the generated response, with only a minor wording change and no substantive restructuring of the reasoning, evidence, or organization.
Supporting reflection	"I used ChatGPT because it was faster. I usually followed the answer because I was not sure how to change it" (S20).

Table A3 *Mixed ChatGPT-use pattern*

Student	Valid sequences	Reflective-evaluative	Product-oriented	Student-level classification
S14	9	5 (55.6%)	4 (44.4%)	Mixed

Mixed use was determined across a student's valid sequences rather than assigned to an individual interaction. The following two sequences from S14 show the shift between reflective-evaluative and product-oriented use.

Sequence A *Reflective-evaluative use*

Element	Record
Week	7
Task focus	Evaluating argument strength
Initial response	Students should be allowed to use AI in writing classes because it can give fast feedback. However, students need to check the feedback because AI does not always understand the purpose of an assignment.
Student prompt	Can you evaluate this argument? Tell me which claim is weak and why. Please do not rewrite it.
ChatGPT response	The main position is understandable, but the first claim is too general. Saying that AI provides "fast feedback" explains efficiency but not why the feedback improves learning. You could clarify what students do with the feedback, such as comparing suggestions with their own ideas or deciding whether revisions are appropriate. The second sentence is stronger because it recognizes a limitation of AI. You could connect the two ideas more clearly by arguing that AI feedback is useful only when students evaluate it rather than accept it automatically.
Follow-up prompt	Is the problem mainly with my first sentence?
ChatGPT follow-up response	Yes. The first sentence focuses on speed, but speed alone is not a strong educational justification. The argument would be stronger if you explained how students actively use AI feedback to improve their writing.
Linked revision	Students should be allowed to use AI in writing classes because its feedback can help them reconsider unclear ideas and identify possible weaknesses in their drafts. However, students need to evaluate the feedback because AI does not always understand the purpose or context of an assignment. AI feedback is therefore most useful when students compare the suggestions with their own ideas before deciding what to revise.
Coding rationale	This sequence was coded as reflective-evaluative because the student requested evaluation of an existing argument, sought clarification about the weakness identified by ChatGPT, and substantially revised the original reasoning rather than reproducing generated text.

Sequence B *Product-oriented use*

Element	Record
Week	9
Task focus	Counterargument and rebuttal
Initial response	Students can use ChatGPT for university assignments because it can help when they do not know how to answer. However, students should not copy everything because they still need to develop their own ideas.
Student prompt	Make a counterargument and rebuttal paragraph about students using ChatGPT for university assignments. About 120 words.
ChatGPT response	Some people argue that university students should be free to use ChatGPT for assignments because AI tools are becoming part of modern education and professional life. From this perspective, restricting ChatGPT may prevent students from developing useful digital skills. However, unrestricted use can also encourage students to rely on generated answers instead of developing their own reasoning. Therefore, universities should not prohibit ChatGPT completely but should establish clear guidelines for its use. Students can use ChatGPT to explore ideas and receive feedback, while the final argument and decisions should remain their own. This approach allows students to develop AI-related skills without weakening their responsibility for academic work.
Linked revision	Some people argue that university students should be free to use ChatGPT for assignments because AI tools are becoming part of modern education and professional life. Restricting ChatGPT may prevent students from developing useful digital skills. However, unrestricted use can encourage students to rely on generated answers instead of developing their own reasoning. Therefore, universities should establish clear guidelines for ChatGPT use. Students can use it to explore ideas and receive feedback, but the final argument should remain their own.
Coding rationale	This sequence was coded as product-oriented because the student moved from a brief initial position to requesting a complete counterargument and rebuttal paragraph, then retained most of the generated argument structure and wording. The revision mainly shortened the ChatGPT response rather than substantively changing its reasoning.

Student-Level Coding Rationale

Across S14's nine valid interaction records, five sequences were coded as reflective-evaluative and four as product-oriented. Neither category reached the provisional 60% criterion for a dominant pattern. S14 was therefore classified as mixed, reflecting variation in ChatGPT use across tasks.

Supporting reflection: "Sometimes I used ChatGPT to get ideas, but when the task was difficult, I asked it to make the paragraph for me first" (S14).

Treatment of Multi-Turn Interaction Records

A task-specific interaction record was coded as one sequence when the initial prompt and subsequent turns addressed the same assigned focal task and led to the same linked revision. For example, S07's initial request to identify weaknesses and the subsequent request for a more appropriate counterargument were treated as one sequence because both exchanges contributed to the revision of the same paragraph. Follow-up turns were used to determine how the student responded to and evaluated the AI output; they were not counted as additional sequences. A separate sequence was recorded only when an interaction concerned a different focal task with its own revision evidence.

Application of the 60% Criterion

The 60% threshold was specified a priori as a provisional analytical criterion for identifying dominant student-level patterns. Students with at least 60% reflective-evaluative sequences were classified as reflective-evaluative dominant, while those with at least 60% product-oriented sequences were classified as product-oriented dominant. Students for whom neither category reached 60% were classified as mixed. The criterion was used to organize behavioural patterns in the present dataset and requires validation across other samples and task settings.