Article

Check for updates

Journal of International Students
Volume 16, Issue 17 (2026), pp. 339-358
ISSN: 2162-3104 (Print), 2166-3750 (Online)
jistudents.org
https://doi.org/10.32674/ggjnen46

A Stage-Specific Affordance Model of ChatGPT Deployment in EFL Writing: Criterion-Level Evidence from a Quasi-Experimental Design

Zahra Zargaran
University of Technology and Applied Sciences, Shinas, Oman

Saif Said Bakhit Aldhahli
University of Technology and Applied Sciences, Shinas, Oman
Corresponding author: Zahra Zargaran; Email: Zahra.zaragaran@utas.edu.om

ABSTRACT: *A stage-specific affordance model of AI-mediated writing instruction predicts that ChatGPT produces qualitatively different learning outcomes depending on the writing stage at which it is deployed. However, this prediction has never been directly tested at the criterion level within a single controlled design. In this study, the model is empirically tested using a quasi-experimental pretest/posttest design. Eighty-four intermediate-level EFL undergraduates at a Gulf-region university were assigned to three intact classes: ChatGPT as a prewriting brainstorming scaffold, ChatGPT as a postdrafting written corrective feedback tool, and a conventional instruction control. Writing was assessed with an institutionally validated four-criterion rubric. One-way ANCOVA with Bonferroni correction revealed a theoretically coherent reversal: the brainstorming group significantly outperformed the feedback group on task response and organization, whereas the feedback group outperformed the brainstorming group on grammatical range and accuracy and vocabulary. Both ChatGPT conditions outperformed the control condition on all the criteria. These findings offer a demonstrative criterion-level validation of a stage-specific affordance model for sequencing ChatGPT in L2 writing.*

Keywords: affordance theory, brainstorming, ChatGPT, EFL writing, quasi-experimental, written corrective feedback

Received: May 25, 2026 | Revised: July 1,2026 | Accepted: July 25, 2026

How to Cite (APA): Zargaran, Z., & Aldhahli, S. S. B. (2026). A stage-specific affordance model of ChatGPT deployment in EFL writing: Criterion-level evidence from a quasi-experimental design. *Journal of International Students*, *16*(17), 339-358. https://doi.org/10.32674/ggjnen46

INTRODUCTION

When does it matter how ChatGPT is deployed, not merely whether it is used? This study proposes and tests a stage-specific affordance model that answers this question at the criterion level of writing assessment. Affordance theory holds that a tool's effects are relational determined not by the tool itself but by what it enables for a particular user at a particular moment in a task (Gibson, 1979; Hutchby, 2001). Applied to ChatGPT in writing instruction, this means that deploying the same tool at the prewriting stage should activate fundamentally different affordances than deploying it at the postdrafting stage, and those differences should produce predictable, criterion-specific learning outcomes. No published study has directly tested this prediction. The present study does so by comparing ChatGPT as a prewriting brainstorming scaffold against ChatGPT as a postdrafting written corrective feedback (WCF) tool within the same experimental design, using the same participants, the same validated rubric, and reporting pairwise criterion-level ANCOVA comparisons between both conditions. The result is a theoretically coherent reversal — each deployment mode outperforming the other on a distinct, predictable subset of writing criteria — that reframes AI integration in L2 writing instruction from a binary choice to a principled, stage-sensitive pedagogical decision.

The literature on ChatGPT in EFL writing has grown rapidly and has reached productive convergence: ChatGPT can improve writing quality (Alharbi, 2023; Huang et al., 2024; Kohnke et al., 2023). What this literature has not resolved and what has far greater pedagogical consequences are whether different ways of deploying ChatGPT produce different patterns of writing development and, specifically, whether those patterns are predictable at the level of individual writing criteria. This distinction matters because writing quality is not a unidimensional construct. The four criteria most commonly assessed in EFL writing—task response, organization, grammatical range and accuracy, and vocabulary—tap qualitatively different competencies, and there is good theoretical reason to expect that AI support at different writing stages should differentially affect these competencies. A brainstorming interaction, for instance, operates at the level of ideational content and structural planning before any text is produced; a feedback interaction operates at the level of grammatical and lexical precision after a draft is complete. Cognitive process writing theory (Flower & Hayes, 1981) maps these directly onto distinct planning and translation/review processes. If the affordance model is correct, criterion-level effects should not merely differ between conditions; they should reverse, with

each condition outperforming the other on a theoretically predictable subset of criteria.

Despite growing interest in AI-assisted EFL writing, existing studies have predominantly examined ChatGPT's effects on overall writing quality without distinguishing between specific deployment functions or reporting criterion-level outcomes (Alharbi, 2023; Kohnke et al., 2023; Yan, 2023). This leaves critical gaps in terms of which deployment function improves which dimension of writing competence; how Gulf Arab EFL learners, an underrepresented context with a distinctive L1 interference profile (Nazari et al., 2021), respond to stage-specific AI support; and whether criterion-level pairwise comparisons between AI deployment conditions can provide the granularity needed for evidence-based pedagogical sequencing. The present study addresses all three gaps within a single quasi-experimental design. Three research questions guided the study:

1. Is there a statistically significant difference in overall writing performance among the three groups following the intervention?
2. Are there significant pairwise differences among all three groups on each of the four writing rubric criteria?
3. What are students' perceptions of ChatGPT as a brainstorming tool and as a feedback tool in terms of usefulness, ease of use, and criterion-specific impact?

LITERATURE REVIEW

Affordance theory and AI-mediated writing

Affordance theory, originally developed by Gibson (1979) in ecological psychology and extended to technology by Hutchby (2001), holds that what a tool affords is not a fixed property of the tool but a relational possibility that emerges between the tool, the user, and the context of use. Applied to ChatGPT in writing instruction, this relational framing predicts that the same tool will activate different possibilities—and produce different outcomes—depending on when and how it is deployed. Writing in a second or foreign language is a cognitively demanding activity requiring the simultaneous management of ideational, organisational, and linguistic resources (Flower & Hayes, 1981). Cognitive process writing theory distributes these demands across sequential stages—prewriting, drafting, revising, and editing—providing a principled framework for mapping which ChatGPT affordances should be most active at each stage and predicting which writing criteria they should most strongly affect.

At the prewriting stage, ChatGPT's key affordances are ideational and structural: its capacity to generate semantically rich content on demand, propose organizational frameworks, and supply genre-appropriate discourse markers without the social costs of peer interaction make it a qualitatively different brainstorming interlocutor (Bai & Hu, 2022). At the postdrafting stage, its key affordances shift to linguistic: its capacity to deliver immediate, criterion-

referenced, multidimensional feedback on grammar and lexical choice represents a qualitatively new form of WCF that differs from both human teacher feedback and earlier automated writing evaluation tools (Ferris, 2004; Guo et al., 2023). This affordance mapping generates the study's central empirical prediction: Brainstorming should primarily improve higher-order criteria (task response and organization), whereas postdrafting feedback should primarily improve lower-order criteria (grammatical range and accuracy and vocabulary).

ChatGPT and AI in EFL Writing Instruction

Research on ChatGPT in EFL writing has grown substantially since the tool's public release (OpenAI, 2022). Early studies reported broadly positive effects on writing quality, with AI-assisted learners outperforming control groups in coherence, grammatical accuracy, and holistic writing scores (Alharbi, 2023; Kohnke et al., 2023). Subsequent work shifted its attention from whether ChatGPT is beneficial to the conditions under which it is most beneficial. Yan (2023) demonstrated that students who received explicit guidance on how to interact with ChatGPT produced significantly better organized essays than those given unrestricted access did. Hong and Shin (2025) reported that structured, goal-directed use promoted deeper task engagement, whereas unconstrained access tended to produce overreliance on AI-generated content. More recent meta-analytic work confirms that effect sizes for AI-assisted writing interventions are moderated substantially by how the tool is deployed, not merely by its presence in the instructional environment (Huang et al., 2024; Zhang & Hyland, 2024).

However, no published study has directly compared distinct ChatGPT deployment functions—brainstorming versus feedback—in the same experimental design, using the same participants and the same validated rubric, and reported pairwise criterion-level comparisons between the two conditions. The present study fills this gap.

ChatGPT as a Brainstorming Tool

The role of prewriting in reducing cognitive load during drafting is well established in both L1 and L2 writing research. Process writing theorists position prewriting scaffolding as essential for enabling writers to focus attentional resources on language production rather than concurrent idea generation (Flower & Hayes, 1981; Zamel, 1982). In EFL contexts, prewriting support has produced measurable improvements in content development and discourse organization (Manchón, 2011; Victori, 1999). Bai and Hu (2022) reported that AI-assisted brainstorming significantly improved idea generation, essay structure, and paragraph coherence among Chinese EFL students, attributing gains to the model's ability to generate semantically rich, topically relevant content that extended the learner's conceptual map of the topic. Yan (2023) reported that AI-assisted outlining improved the organizational quality of EFL comparison essays, particularly for students who struggled with identifying and sequencing appropriate discourse structures.

The comparison-and-contrast genre used in the present study has a particular affinity with AI-assisted brainstorming: its demands for point-by-point structure, discourse marker management, and parallel development map directly onto what ChatGPT scaffolds at the planning stage (Hyland, 2004). Unlike peer brainstorming, ChatGPT offers an ideational interlocutor that is available on demand, requires no social negotiation, and can generate genre-specific structural scaffolding instantaneously—a qualitatively different resource for Arabic-speaking EFL learners whose L1 discourse preferences may not align with the linear structure expected in academic English (Nazari et al., 2021).

ChatGPT as a Written Corrective Feedback Tool

The role of WCF in supporting L2 writing accuracy has been extensively studied and meta-analytically confirmed. Ferris (2004) and Bitchener and Knoch (2010) reported that direct, focused WCF on grammatical errors leads to sustained accuracy gains when learners actively revise their texts in response to the feedback received. Kang and Han's (2015) meta-analysis confirmed a moderate to large effect of WCF on L2 writing accuracy, with the strongest gains observed for grammatical and lexical features. ChatGPT has emerged as a novel mechanism for delivering such feedback at scale. Guo et al. (2023) reported ChatGPT to be comparably effective for improving sentence-level accuracy and vocabulary while outperforming teacher feedback on immediacy and specificity. Cavaleri and Dianati (2023) reported that compared with peer feedback, ChatGPT feedback prompted substantially more lexical and grammatical revision. Zhang and Liu (2026) reported that compared with generic AI feedback, rubric-aligned ChatGPT prompting resulted in significantly more targeted accuracy revisions, underscoring the importance of structured prompt design.

A theoretically important limitation has been consistently identified: ChatGPT's discourse-level commentary is less precise than its sentence-level feedback (Guo et al., 2023; Huang et al., 2024). This directly predicts stronger feedback-condition gains on accuracy criteria than on higher-order organisational criteria—a prediction tested at the pairwise level in the present study and central to the stage-specific affordance model being advanced here.

Gulf Arab EFL Learners

Gulf Arab EFL learners present a distinctive L1 interference profile—including Arabic verb-subject-object word order, nominal sentence structures, and coordinate clause-chaining preferences—that diverges markedly from academic English discourse structure (Hyland, 2004; Nazari et al., 2021). Al-Shabandar et al. (2022) documented strongly positive attitudes toward AI feedback tools among Gulf-region students, making this an important and underrepresented context for investigating deployment-specific affordance effects. Perceived usefulness and ease of use have likewise been identified as central determinants of international students' adoption of generative AI tools (Ittefaq et al., 2025), and recent evidence from a Gulf-region English–medium university shows that structured, protocol-

driven integration of generative AI into writing courses yields large and more equitable learning gains (Nelson, 2026). The present study extends this work by providing the first controlled, criterion-level comparison of distinct ChatGPT deployment modes within this learner population.

METHODOLOGY

Research Design

This study employed a quasi-experimental pretest/posttest design with three intact classes. A triangulated quantitative design combined writing score data from ANCOVAs with descriptive perception questionnaire data and supplementary ChatGPT interaction log analysis to examine performance outcomes, student perceptions, and process-level engagement. This study is theoretically grounded in Flower and Hayes's (1981) cognitive process model of writing, which provides a conceptual framework for distinguishing the prewriting and postdrafting phases as sites of differential AI support.

A key constraint of intact-class designs warrants explicit acknowledgment: teacher, classroom dynamics, and peer effects are confounded with treatment assignment. ANCOVA controls for preexisting ability differences but cannot eliminate between-classroom confounds. All causal language is therefore provisional, and between-group ANCOVA comparisons — not within-group gains — provide the basis for inferential conclusions. This limitation and its implications are addressed in the Discussion section.

Participants

All participants were intermediate-level EFL undergraduates enrolled in a compulsory general English writing course at a Gulf-region university [blinded for review]. The three intact classes were allocated 30 seats each by the institutional registrar, resulting in equal enrollment across all sections prior to the study. Of the 90 students who were initially enrolled (30 per class), six withdrew before the intervention commenced, resulting in a final analytic sample of 84 participants: Group A (ChatGPT as a brainstorming tool, n = 28), Group B (ChatGPT as a feedback tool, n = 28), and Group C (conventional instruction control, n = 28). All participants were Arabic-speaking learners who were placed at the same institutional proficiency level through the university's standardized placement procedure. A pretest one-way ANOVA confirmed that there was no statistically significant difference in initial writing ability across the three groups, $F(2, 81) = 0.19$, $p = .83$, confirming baseline equivalence.

Ethical approval was obtained from the institutional Research Ethics Committee (blinded for review) prior to data collection, and all participants provided written informed consent. Participation was voluntary, with no academic penalty for withdrawal.

Instruments

Writing Rubric

Writing performance was assessed using the institutionally validated general English writing rubric for intermediate students (GE2), which evaluates four equally weighted criteria—task response, organization, grammatical range and accuracy, and vocabulary—each scored 1–5 (maximum 20). Two trained writing instructors and the researcher independently double-marked all the scripts; the interrater reliability was strong (Cohen's $\kappa = .84$), and discrepancies were resolved through structured discussion. The rubric was validated for use across all general English writing courses at the institution through a faculty-wide calibration process.

Perception Questionnaire

A 20-item Likert-scale questionnaire (1 = strongly disagree to 5 = strongly agree) was administered postintervention to Groups A and B to measure perceived usefulness, ease of use, criterion-specific impact, and overall satisfaction. Items were adapted from Davis' (1989) Technology Acceptance Model, a framework recently extended to international students' adoption of generative AI (Ittefaq et al., 2025) and validated by a panel of five experts (three EFL writing specialists and two educational technology researchers) using a four-point content validity index (CVI). Items with a CVI below .80 were revised or replaced before finalization. The internal consistency was acceptable (Cronbach's $\alpha = .87$).

ChatGPT Interaction Logs

All student–ChatGPT interactions were logged via the shared institutional ChatGPT account during class sessions, producing a total of 672 session logs (336 per treatment group across four tasks). Logs were coded by two independent raters for exchange type (ideation/structure elicitation for Group A; error identification/revision request for Group B), uptake evidence, and protocol fidelity. Interrater agreement was high ($\kappa = .89$); discrepancies were resolved through discussion.

Prompt Standardization

To ensure the comparability of student–ChatGPT interactions within each condition, standardized prompt templates were developed, validated, and distributed during the week 1 training session. Group A used three sequential prompt types before drafting (topic exploration, outline generation, and discourse marker elicitation), and Group B used two sequential prompt types after completing an independent draft (criterion-referenced feedback and revision confirmation). Full templates are provided in Appendix A. Promotion compliance was verified through biweekly log review; deviations were flagged and addressed

with participants. Protocol fidelity was high: fewer than 1% of logged sessions showed evidence of off-protocol use.

Procedure

The intervention spanned eight weeks. All groups completed a 45-minute comparison-and-contrast writing pretest under standardized conditions in Week 1; Groups A and B also received structured ChatGPT training using the prompt templates during Week 1. Group A used ChatGPT exclusively during prewriting and ceased all AI interactions before it was drafted. Group B completed a full independent draft before being submitted to ChatGPT for rubric-aligned feedback and revision. Group C received standard teacher-led writing instruction with no ChatGPT exposure.

Over weeks 2–7, all groups completed four writing tasks in 90-minute sessions; total writing time was held equal across groups. Trained instructors (not the researcher) delivered all the sessions to minimize researcher bias. A posttest was administered in Week 8 under the same standardized conditions as the pretest, and the perception questionnaire was administered to Groups A and B immediately afterwards.

Data Analysis

All the required parametric assumptions were tested and confirmed prior to analysis (see Table 1). One-way ANCOVA with pretest scores as the covariate was used to compare adjusted posttest total scores across the three groups, followed by Bonferroni-corrected pairwise post hoc comparisons for all three pairs (A vs. B, A vs. C, B vs. C). Separate criterion-level ANCOVAs were run for each of the four rubric criteria, each followed by identical Bonferroni-corrected pairwise comparisons. Treatment effect sizes are reported as partial η^2. Within-group gains are summarized descriptively using Cohen's *d*, which is calculated as M_{diff}/SD_{diff} following Cohen (1988). ChatGPT interaction log data were analyzed using descriptive statistics and interrater coded content analysis. The questionnaire data were analyzed descriptively (means and standard deviations) by item clustering. The use of ANCOVA was considered appropriate because the study employed a quasi-experimental pretest/posttest design with intact classes rather than random assignment at the individual level. By statistically controlling for baseline differences in pretest writing performance, ANCOVA provides a more precise estimate of the treatment effect on posttest outcomes while reducing error variance and improving statistical power. This analytical approach is widely recommended for educational intervention studies in which preexisting group differences cannot be eliminated through randomization, provided that the underlying assumptions of ANCOVA are satisfied. Accordingly, all the required assumptions were examined and confirmed prior to analysis. All analyses were conducted in IBM SPSS v.26 with $\alpha = .05$.

Ethical Considerations

Ethical principles governing research involving human participants were observed throughout the study. Prior to data collection, ethical approval to conduct the study was obtained from the appropriate institutional authority in accordance with the institution's research ethics procedures. Participation was entirely voluntary, and all the students were informed of the study's purpose, procedures, and their right to withdraw at any stage without academic penalty. Written informed consent was obtained from all participants before the intervention began. To protect participants' privacy, all the data were anonymized prior to analysis, and no personally identifiable information was collected or reported. Furthermore, ChatGPT was used solely as part of the planned instructional intervention under the supervision of the course instructor, and participants were instructed to use the tool only within the prescribed pedagogical framework to ensure academic integrity and equitable participation across all study conditions.

RESULTS

Preliminary Assumption Checks

All the parametric assumptions were confirmed prior to analysis (Table 1). Levene's test confirmed the homogeneity of variance, $F(2, 81) = 0.24$, $p = .79$; the Group × Pretest interaction was nonsignificant, $F(2, 78) = 0.61$, $p = .55$, confirming the homogeneity of regression slopes; Shapiro–Wilk tests on residuals were nonsignificant (all ps > .05); and pretest equivalence confirmed covariate independence, $F(2, 81) = 0.19$, $p = .83$.

Table 1: Parametric Assumption Check Results

Assumption Test	Statistic	*df*	*p*
Levene's Test (Homogeneity of Variance)	$F = 0.24$	2, 81	.79
Homogeneity of Regression Slopes (Group × Pretest)	$F = 0.61$	2, 78	.55
Shapiro–Wilk — Group A Residuals	$W = 0.977$	27	.782
Shapiro–Wilk — Group B Residuals	$W = 0.981$	27	.851
Shapiro–Wilk — Group C Residuals	$W = 0.974$	27	.701
Covariate Independence (Pretest ANOVA)	$F = 0.19$	2, 81	.83

ChatGPT Interaction Log Analysis

Interaction log analysis provided process-level evidence of differential engagement patterns. Group A students produced a mean of 8.3 ChatGPT exchanges per session (*SD* = 1.4), 94% of which were classified as ideation or structure-elicitation queries. Postwriting log review confirmed that 91% of Group A participants incorporated ChatGPT-generated outline elements into their submitted drafts. Group B students produced a mean of 6.1 exchanges per session (*SD* = 1.7), of which 89% were classified as error-identification or revision-request queries. Revision uptake analysis confirmed that 87% of Group B participants demonstrably incorporated at least one AI-identified linguistic suggestion into their final drafts. Protocol fidelity was high: only 4 of 672 logged sessions (< 1%) showed evidence of off-protocol use, confirming that the observed differences in outcomes reflect genuine treatment variation rather than procedural drift.

Overall Writing Performance

Pretest means were statistically equivalent across groups (range: 10.25–10.43; see Table 2). Both treatment groups produced substantially larger gains than the control ($d \approx 1.57$–1.70 vs. $d = 0.42$), although Groups A and B achieved virtually identical adjusted posttest totals (13.12 vs. 13.23), indicating that neither ChatGPT condition was globally superior to the other at the total score level.

Table 2: Descriptive Statistics for Pretest and Posttest Total Scores by Group

Group	Pre *M*	Pre *SD*	Post *M*	Post *SD*	Adj. *M*	Gain	Cohen's *d*
A — Brainstorming (*n* = 28)	10.43	1.89	13.14	1.72	13.12	2.71	1.57
B — Feedback (*n* = 28)	10.36	1.92	13.21	1.68	13.23	2.85	1.70
C — Control (*n* = 28)	10.25	1.88	11.04	1.90	11.08	0.79	0.42

Note. Cohen's *d* was calculated as M_{diff}/SD_{diff} (Cohen, 1988). Adj. *M* = ANCOVA-adjusted posttest mean.

ANCOVA and Pairwise Comparisons: Overall Score

ANCOVA revealed a significant main effect of group, $F(2, 80) = 23.41$, $p < .001$, partial $\eta^2 = .37$ (large effect). Bonferroni-corrected comparisons confirmed significant differences for A vs. C ($p < .001$) and B vs. C ($p < .001$) but not A vs. B (mean difference = 0.11, $p = .91$): the two treatment conditions produced equivalent overall gains. The results are presented in Table 3.

Table 3: ANCOVA Results and Pairwise Comparisons for the Total Writing Score

Source	*SS*	*df*	*MS*	*F*	*p*	partial η^2
Pretest (Covariate)	28.41	1	28.41	42.17	< .001	.34
Group (Treatment)	31.56	2	15.78	23.41	< .001	.37
Error	54.37	80	0.68			
Corrected Total	114.34	83				

Criterion-level Analysis: Central Findings

The criterion-level analysis revealed the theoretically predicted reversal pattern (Table 4). With respect to task response and organization, Group A (brainstorming) significantly outperformed Group B (feedback; $p = .03$ and $p = .02$, respectively) and Group C (both ps < .001), whereas Group B outperformed Group C (both ps < .001). The pattern reversed precisely for Grammatical Range and Accuracy and Vocabulary: Group B significantly outperformed Group A ($p = .04$ for both criteria) and Group C (both ps < .001), whereas Group A still outperformed Group C (both ps < .001). All omnibus F statistics were significant (ps < .001, partial $\eta^2 = .29$–.34, all large effects). Systematic reversal—each ChatGPT condition significantly outperforms the other on a theoretically predictable and distinct subset of criteria—conforms the central empirical findings of this study and the primary evidence for the stage-specific affordance model.

Table 4: Criterion-Level Adjusted Posttest Means and Pairwise Comparisons (Bonferroni-Corrected)

Group	Task Response	Organization	GRA	Vocabulary
A — Brainstorming	3.46	3.57	3.04	3.07
B — Feedback	3.07	3.00	3.54	3.61
C — Control	2.71	2.68	2.79	2.86
$F(2, 80)$	16.28***	17.41***	18.73***	21.04***
partial η^2	.29	.30	.32	.34
A vs. B	$p = .03$*	$p = .02$*	$p = .04$*	$p = .04$*
A vs. C	$p < .001$***	$p < .001$***	$p < .001$***	$p < .001$***
B vs. C	$p < .001$***	$p < .001$***	$p < .001$***	$p < .001$***

Note. GRA = Grammatical Range and Accuracy. Adjusted means derived from one-way ANCOVA with pretest score as a covariate. *$p < .05$. ***$p < .001$. Direction of superiority: A > B on Task Response and Organization; B > A on GRA and Vocabulary.

Student Perceptions

Both groups reported high overall satisfaction with ChatGPT. Critically, the perception ratings mirrored objective performance: Group A rated ChatGPT's

impact on Task Response (M = 4.42, SD = 0.51) and Organization (M = 4.38, SD = 0.53) substantially higher than Group B (M = 3.74, SD = 0.68; M = 3.61, SD = 0.71), whereas Group B rated its impact on Grammatical Range and Accuracy (M = 4.47, SD = 0.52) and Vocabulary (M = 4.51, SD = 0.49) substantially greater than Group A (M = 3.56, SD = 0.69; M = 3.48, SD = 0.72). This convergence between objective gains and subjective perceptions provides concurrent validity for the quantitative findings. The full questionnaire descriptive statistics are presented in Table 5.

Table 5: Questionnaire Descriptive Statistics by Group (Scale 1–5)

Item Cluster	Group A — Brainstorming *M* (*SD*)	Group B — Feedback *M* (*SD*)
Perceived Usefulness	4.31 (0.54)	4.18 (0.61)
Ease of Use	4.12 (0.58)	4.09 (0.63)
Impact on Task Response	4.42 (0.51)	3.74 (0.68)
Impact on Organization	4.38 (0.53)	3.61 (0.71)
Impact on Grammatical Range & Accuracy	3.56 (0.69)	4.47 (0.52)
Impact on Vocabulary	3.48 (0.72)	4.51 (0.49)
Overall Satisfaction	4.27 (0.55)	4.33 (0.57)

DISCUSSION

Stage-Specific Affordance Model: Empirical Validation

The central finding of this study—a systematic, theoretically predicted reversal of criterion-level effects across the two ChatGPT deployment conditions—conforms the first direct empirical test of a stage-specific affordance model of AI-mediated L2 writing instruction. From an affordance perspective (Gibson, 1979; Hutchby, 2001), different deployment contexts activate different relational possibilities between the tool and the learner. The pattern of results maps precisely onto Flower and Hayes's (1981) distinction between planning processes and translation/reviewing processes: Brainstorming gains are localized to higher-order criteria (task response, organization), whereas feedback gains are localized to lower-order criteria (grammatical range and accuracy, vocabulary). ChatGPT operates as a stage-specific cognitive prosthesis, amplifying the processes most active at the writing stage at which it is deployed.

These findings advance literature in a theoretically consequential way. Prior research has shown that ChatGPT can improve EFL writing quality (Alharbi, 2023; Huang et al., 2024) and that structured use is more effective than unstructured access (Yan, 2023). The present study goes further by demonstrating that the mode of deployment not only moderates overall effectiveness but also produces predictable, criterion-differentiated outcomes—a result that transforms the question from 'does AI help?' to 'which AI function helps which dimension of writing, and why?

Brainstorming and Higher-Order Writing Criteria

Group A's superior gains in task response and organization are consistent with cognitive process writing theory: prewriting planning reduces the cognitive load by enabling writers to enter the composing phase with a richer mental representation of the task (Flower & Hayes, 1981). Interaction log analysis corroborated this mechanism; 91% of Group A participants incorporated ChatGPT-generated outline elements into their drafts, confirming that AI-mediated brainstorming translated into structural scaffolding rather than surface protocol compliance.

Unlike peer brainstorming, ChatGPT offers an ideational interlocutor that is available on demand, requires no social negotiation, and can generate genre-specific structural scaffolding instantaneously. This is a qualitatively different resource for Arabic-speaking EFL learners whose L1 discourse preferences—coordinate clause-chaining and nominal sentence structures—may not align with the linear, point-by-point structure of comparison-and-contrast academic writing in English (Hyland, 2004; Nazari et al., 2021). The brainstorming group's relatively modest gains in linguistic criteria are theoretically coherent: prewriting interaction operates at the level of ideas and structure, not at the sentence level, and students cease AI interaction before they draft, providing no opportunity for linguistic refinement.

Feedback and Lower-Order Writing Criteria

Group B's superior gains in Grammatical Range and Accuracy and Vocabulary extend the established WCF findings to AI-generated, rubric-aligned feedback delivered at scale. The feedback revision cycle — independent drafting, ChatGPT criterion-referenced feedback, and revision — created repeated opportunities for focused attention to form and lexical precision across the four tasks, which is consistent with the conditions that Ferris (2004) and Kang and Han (2015) identify as most conducive to accuracy development. Log analysis confirmed 87% active uptake of AI-identified suggestions, indicating genuine engagement rather than passive receipt.

The feedback group's lower gains on task response and organization are explicable by timing: structural weaknesses were already embedded in drafts before ChatGPT was consulted, and ChatGPT's discourse-level feedback is consistently reported to be less precise than its sentence-level commentary (Guo et al., 2023; Huang et al., 2024). The affordance model accounts for this asymmetry: postdrafting interaction activates linguistic affordances but cannot retroactively restructure ideational and organisational planning that has already been completed.

Both ChatGPT Modes Outperform Conventional Instruction

Both ChatGPT conditions produced significantly greater overall writing gains than conventional instruction did, with large within-group effect sizes, which is

consistent with meta-analytic evidence for structured AI tool use in language learning (Huang et al., 2024; Zhang & Hyland, 2024). Two alternative explanations cannot be fully excluded: novelty-induced motivation and additional cognitive processing time in ChatGPT conditions. Active digital control conditions with equal time-on-task are needed to isolate AI-specific effects from novelty effects, and future research should prioritize this design refinement.

Beyond the immediate EFL classroom, these findings have broader implications for English–medium higher education institutions that increasingly serve linguistically and culturally diverse student populations. International and multilingual students often face simultaneous challenges related to academic writing, disciplinary discourse conventions, and adaptation to new educational expectations. Within such contexts, the pedagogically strategic deployment of generative AI has the potential to provide timely and individualized support while complementing, rather than replacing, instructor guidance. The stage-specific affordance model proposed in this study suggests that AI tools may be most effective when they are integrated systematically into writing instruction according to learners' developmental needs and the cognitive demands of different writing stages. Such an approach may help universities promote equitable access to academic writing support while fostering the responsible and pedagogically informed use of generative AI in international higher education.

Pedagogical Implications

The criterion-level pairwise findings support a sequenced instructional approach: deploying ChatGPT as a brainstorming scaffold in early writing tasks to activate ideational and organizational affordances and then introducing rubric-aligned feedback in later tasks to activate linguistic affordances is likely to yield more comprehensive criterion coverage than either mode alone. The prompt standardization framework (Appendix A) provides a directly replicable template for such a sequenced implementation. Both modes require structured protocols and explicit training: unguided ChatGPT use risks superficial engagement, overreliance on AI-generated content, and academic integrity violations (Bannister et al., 2024; Chiu, 2024; Lo, 2023). Such risks are especially salient for international and second-language students, for whom institutional policies on generative AI use remain inconsistent across higher education systems (Bannister et al., 2024). Embedding AI within an established pedagogical progression—rather than treating it as an optional supplement—may be key to realizing its developmental potential.

Limitations and Future Research

Several limitations constrain interpretation and point toward future research priorities. The quasi-experimental intact-class design is the most significant constraint: with a single class per condition, teacher and peer effects cannot be disentangled from treatment effects, and the Hawthorne effect cannot be excluded. ANCOVA adjusts for preexisting ability differences but does not

resolve between-classroom confounds. Future research should employ true random assignments at the individual level where ethically and institutionally feasible or adopt multiclass designs with multiple teachers per condition.

The study is further limited to a single institution, a single writing genre (comparison-and-contrast), and an eight-week intervention window. Whether the criterion-level reversal pattern replicates across genres, proficiency levels, and instructional contexts remains an important empirical question. Interaction log analysis was limited to exchange frequency and type; analysis of the quality of AI responses and the depth of student engagement with those responses would provide richer mechanistic evidence.

Future research priorities include (a) longitudinal designs to assess whether criterion-level gains are durable beyond the intervention period; (b) studies of sequenced ChatGPT deployment—brainstorming in early tasks and feedback in later tasks—to test whether the affordance effects are additive; (c) think-aloud protocols to illuminate the cognitive processes mediating criterion-level gains; and (d) replication in other L1 contexts to test the boundary conditions of the stage-specific affordance model.

CONCLUSION

This study proposed and tested a stage-specific affordance model of AI-mediated L2 writing instruction, providing the first criterion-level empirical evidence that distinct ChatGPT deployment modes produce theoretically predictable, reversed patterns of writing development. Criterion-level ANCOVA with Bonferroni-corrected pairwise comparisons confirmed that the brainstorming condition produced significantly superior gains in terms of task response and organization, whereas the feedback condition produced significantly superior gains in terms of gray range and accuracy and vocabulary; each condition was superior in terms of precisely the criteria its affordance profile predicted. Both ChatGPT modes significantly outperformed conventional instruction on all the criteria, and the perception data and interaction logs independently corroborated this differential pattern.

These findings reframe AI integration in L2 writing instruction from a binary question of whether to use ChatGPT to a principled question of which affordances to activate at which writing stage. The stage-specific affordance model, which is grounded in Flower and Hayes's (1981) cognitive process framework and Gibson's (1979) affordance theory, provides CALL researchers and L2 writing practitioners with a theoretically grounded basis for sequencing AI tool use: brainstorming first for ideational and organizational affordances and feedback later for linguistic affordances. Future research should test whether these effects are additive when sequenced within the same instructional program and replicate the criterion-level reversal across genres, contexts, and learner populations.

Declaration of competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Use of AI-Assisted Technologies

No generative AI tools were used to write this manuscript, produce graphical elements, or collect and analyze the data. ChatGPT was used exclusively by the student participants as part of the instructional intervention, as described in the Methodology section.

REFERENCES

Alharbi, M. A. (2023). Exploring the potential of ChatGPT as an EFL writing assistant: Benefits and challenges. *English as a Foreign Language International Journal, 27*(1), 1–18. https://doi.org/10.24205/03276716.2023.1

Al-Shabandar, R., Hussain, A., & Liatsis, P. (2022). Artificial intelligence in EFL writing: Attitudes and perceptions of Gulf-region learners. *Computer Assisted Language Learning, 35*(9), 2190–2215. https://doi.org/10.1080/09588221.2021.1970756

Bai, L., & Hu, G. (2022). AI-assisted brainstorming and its effects on EFL learners' argumentative writing. *System, 109*, Article 102864. https://doi.org/10.1016/j.system.2022.102864

Bannister, P., Alcalde Peñalver, E., & Santamaría Urbieta, A. (2024). International students and generative artificial intelligence: A cross-cultural exploration of HE academic integrity policy. *Journal of International Students, 14*(3), 149–170. https://doi.org/10.32674/jis.v14i3.6277

Bitchener, J., & Knoch, U. (2010). Raising the linguistic accuracy level of advanced L2 writers with written corrective feedback. *Journal of Second Language Writing, 19*(4), 207–217. https://doi.org/10.1016/j.jslw.2010.10.002

Cavaleri, M., & Dianati, S. (2023). You want me to check your grammar again? The usefulness of ChatGPT as a grammar checker and language tutor for EFL learners. *Journal of China Computer-Assisted Language Learning, 3*(1), 137–154. https://doi.org/10.1515/jccall-2023-0009

Chiu, T. K. F. (2024). Future research recommendations for transforming higher education with generative AI. *Computers and Education: Artificial Intelligence, 6*, Article 100197. https://doi.org/10.1016/j.caeai.2023.100197

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

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 13*(3), 319–340. https://doi.org/10.2307/249008

Ferris, D. R. (2004). The "grammar correction" debate in L2 writing: Where are we, and where do we go from here? *Journal of Second Language Writing, 13*(1), 49–62. https://doi.org/10.1016/j.jslw.2004.04.005

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication, 32*(4), 365–387. https://doi.org/10.2307/356600

Gibson, J. J. (1979). *The ecological approach to visual perception*. Houghton Mifflin.

Guo, K., Wang, J., & Chu, S. K. W. (2023). Using ChatGPT to provide formative feedback on EFL writing: Comparing AI and teacher feedback. *Language Learning & Technology, 27*(2), 1–23. https://www.lltjournal.org/item/3676

Hong, S., & Shin, Y. K. (2025). Effects of three levels of AI integration on second language academic writing: Evaluating restricted, guided, and free use of ChatGPT. *System, 134*, Article 103820. https://doi.org/10.1016/j.system.2025.103820

Huang, X., Zou, D., Cheng, G., & Xie, H. (2024). A systematic review and meta-analysis of AI-assisted language learning: Moderating effects of tool type, learning focus, and study design. *Computers & Education, 199*, Article 104967. https://doi.org/10.1016/j.compedu.2023.104967

Hutchby, I. (2001). Technologies, texts and affordances. *Sociology, 35*(2), 441–456. https://doi.org/10.1177/S0038038501000219

Hyland, K. (2004). *Genre and second language writing*. University of Michigan Press.

Ittefaq, M., Zain, A., Arif, R., Ahmad, T., Khan, L., & Seo, H. (2025). Factors influencing international students' adoption of generative artificial intelligence: The mediating role of perceived values and attitudes. *Journal of International Students, 15*(7), 127–156. https://doi.org/10.32674/fnwdpn48

Kang, E., & Han, Z. (2015). The efficacy of written corrective feedback in improving L2 written accuracy: A meta-analysis. *Modern Language Journal, 99*(1), 1–18. https://doi.org/10.1111/modl.12189

Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for EFL/ESL writing: Affordances, limitations, and the way forward. *RELC Journal, 54*(2), 537–550. https://doi.org/10.1177/00336882231162336

Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences, 13*(4), Article 410. https://doi.org/10.3390/educsci13040410

Manchón, R. M. (Ed.). (2011). *Learning-to-write and writing-to-learn in an additional language*. John Benjamins.

Nazari, N., Shabbir, M. S., & Setiawan, R. (2021). Application of artificial intelligence powered digital writing assistant in higher education: Randomized controlled trial. *Heliyon, 7*(5), Article e07014. https://doi.org/10.1016/j.heliyon.2021.e07014

Nelson, E. C. (2026). Structured AI workflows in Saudi EMI research writing: Performance gains, equity, and academic integrity across three semesters. *Journal of International Students, 16*(15), 81–102. https://doi.org/10.32674/e7jdhc67

OpenAI. (2022). *Introducing ChatGPT*. https://openai.com/blog/chatgpt
Victori, M. (1999). An analysis of writing knowledge in EFL composing: A case study of two effective and two less effective writers. *System, 27*(4), 537–555. https://doi.org/10.1016/S0346-251X(99)00049-4
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
Zamel, V. (1982). Writing: The process of discovering meaning. *TESOL Quarterly, 16*(2), 195–209. https://doi.org/10.2307/3586792
Zhang, Y., & Liu, Y. (2026). Designing ChatGPT-mediated feedback activities in EFL writing: A design-based study of the dialogic feedback triangle. *Assessment & Evaluation in Higher Education, 51*(1), 137–160. https://doi.org/10.1080/02602938.2025.2571846
Zhang, Z. (V.), & Hyland, K. (2024). The role of digital literacy in student engagement with automated writing evaluation (AWE) feedback on second language writing. *Computer Assisted Language Learning*. Advance online publication. https://doi.org/10.1080/09588221.2023.2256815

APPENDIX A

Standardized ChatGPT Prompt Templates

Group A—Brainstorming Prompts (Prewriting Phase)

Prompt 1—Topic Exploration: "I am writing a comparison-and-contrast essay about [TOPIC]. Please suggest the five most important points of comparison or contrast I should consider, and briefly explain why each point is significant for an academic essay."
Prompt 2—Outline Generation: "On the basis of the comparison points we discussed, please suggest a point-by-point outline for my essay. Include a suggested topic sentence for each body paragraph and note which transition words or phrases would be appropriate at each stage."
Prompt 3—Discourse Marker Elicitation: "What academic transition words and phrases would be most appropriate for connecting the sections of a comparison-and-contrast essay? Please provide examples of how each might be used in a sentence."
Note: Students were instructed to cease all AI interactions after completing Prompt 3 and before beginning to draft. Participants were reminded that any text generated by ChatGPT must not be copied into their essays; ChatGPT output was to be used only as a planning resource.

Group B—Feedback Prompts (Postdrafting Phase)

Prompt 1—Criterion-Referenced Feedback: "I have written an essay, which I will paste below. Please evaluate it against the following four criteria: (1) Task response — have I fully addressed the essay question and developed clear

comparison points? (2) Organization—is the essay clearly structured with logical progression and appropriate transitions? (3) Grammatical Range and Accuracy — are there grammatical errors, and is there sufficient variety in my sentence structures? (4) Vocabulary—is my word choice accurate, appropriate in the register, and sufficiently varied? Please provide specific suggestions for improvement under each criterion." [Student's essay pasted here]
Prompt 2—Revision Confirmation: "I have made the following revisions on the basis of your feedback: [student lists revisions]. Have I adequately addressed the issues you identified? Are there any remaining weaknesses I should correct before submitting?"
Note: Students were instructed to complete a full independent draft before initiating any ChatGPT interaction. The rubric-aligned feedback prompt was provided to all Group B participants as a printed card to ensure consistency of use across sessions.

Author Bios

ZAHRA ZARGARAN, PhD, is an Assistant Professor of Applied Linguistics and ELT at the University of Technology and Applied Sciences, Shinas, Oman, with more than 19 years of university-level teaching experience across Iran and Oman. Her research spans teacher cognition, language assessment, AI-integrated pedagogy, and written feedback, with publications in international and Scopus-indexed journals. She is a certified trainer by the British Council, Cambridge University, and IDP Australia. Dr. Zargaran is committed to advancing language education through research, innovation, and professional development. Email: Zahra.zaragaran@utas.edu.om

SAIF SAID BAKHIT ALDHAHLI, MA, is a lecturer at the University of Technology and Applied Sciences, Shinas, Oman, with an MA in Applied Linguistics from the University of Bedfordshire. His research resides at the intersection of second language acquisition (SLA) and curriculum design, with a particular focus on fostering learner autonomy through innovative teaching methods. He is dedicated to refining assessment and feedback frameworks to better support student growth, alongside his specialized interest in Phonetic Engineering and the systematic development of oral proficiency. Email: Saif.Aldhahli@utas.edu.om