Journal of English Language Teaching and Applied Linguistics
ISSN: 2707-756X
DOI: 10.32996/jeltal
Journal Homepage: www.al-kindipublisher.com/index.php/jeltal

The Impact of Generative and Corrective AI Tools (ChatGPT and Grammarly) on Developing Academic Writing Proficiency Among Saudi University EFL Students: A Mixed-Methods Investigation

Hiba Ahmed Balal Ahmed

Department of English Language and Literature, College of Languages and Humanities, Qassim University

Corresponding Author: Hiba Ahmed Balal Ahmed**, E-mail:** hs.ahmed@qu.edu.sa

ABSTRACT

While generative and corrective artificial intelligence (AI) tools are increasingly adopted by university learners, empirical evidence systematically comparing their distinct pedagogical impacts on academic writing sub-skills remains limited. Grounded in Vygotsky's Sociocultural Theory (Zone of Proximal Development) and Davis's Technology Acceptance Model (TAM), this study investigates how generative AI (ChatGPT) versus automated corrective AI (Grammarly) affects academic writing proficiency among Saudi university English as a Foreign Language (EFL) students. Using an explanatory sequential mixed-methods quasi-experimental design, 99 undergraduate English majors at a Saudi public university were assigned across three conditions over 12 weeks: Experimental Group 1 (EG1: ChatGPT, n = 34), Experimental Group 2 (EG2: Grammarly, n = 33), and a Control Group (CG: traditional instructor feedback, n = 32). Quantitative data included pre- and post-intervention argumentative essays evaluated via a 20-point Analytic Writing Rubric and a TAM perception survey, analyzed using ANCOVA and MANOVA. Qualitative data were gathered via semi-structured interviews (n = 15) and analyzed thematically. One-way ANCOVA revealed that both EG1 ($M_{\text{adj}} = 16.21$) and EG2 ($M_{\text{adj}} = 15.05$) significantly outperformed CG ($M_{\text{adj}} = 13.13$, $\eta_p^2 = 0.449$), with EG1 achieving higher overall gains than EG2 ($p = 0.003$). MANOVA demonstrated functional divergence: ChatGPT yielded superior gains in macro-level content/organization and coherence/cohesion ($p < 0.001$), whereas Grammarly produced the highest grammatical accuracy ($p = 0.042$). Thematic findings underscored the value of ethical prompt training, metalinguistic reflection, and a sequenced hybrid workflow. This study provides granular sub-skill benchmarks and an actionable, ethically governed hybrid AI writing framework tailored to Saudi higher education under Vision 2030.

KEYWORDS

Academic Writing; Artificial Intelligence; ChatGPT; Grammarly; Saudi EFL Learners; Mixed-Methods Research; Zone of Proximal Development.

ARTICLE INFORMATION

ACCEPTED: 20 October 2026 **PUBLISHED:** 20 November 2026 **DOI:** 10.32996/jeltal.2026.8.12.1

1. Introduction

1.1 Background of the Study

In the contemporary landscape of higher education, English as a Foreign Language (EFL) writing instruction is undergoing a rapid paradigm shift propelled by advancements in artificial intelligence (AI). Within Saudi universities, mastering academic writing in English is a core requirement for undergraduate and postgraduate success, directly influencing students' scholarly communication, critical thinking, and readiness for globalized labor markets under Saudi Vision 2030. However, producing coherent, grammatically accurate, and rhetorically appropriate academic prose remains one of the most cognitively demanding skills for Saudi EFL learners (Alshammari et al., 2022).

Traditionally, tertiary writing pedagogy has relied heavily on instructor-led written corrective feedback (WCF). While valuable, high student-to-faculty ratios and heavy instructional workloads frequently delay feedback delivery, diminishing its pedagogical efficacy during the iterative drafting and revision stages (Hyland & Hyland, 2019). Recent breakthroughs in natural language processing have introduced two distinct categories of AI writing assistants into university classrooms: generative conversational AI (exemplified by ChatGPT) and automated corrective feedback (ACF) tools (exemplified by Grammarly). Whereas Grammarly specializes in real-

Copyright: © 2026 the Author(s). This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) 4.0 license (https://creativecommons.org/licenses/by/4.0/). Published by Al-Kindi Centre for Research and Development, London, United Kingdom.

time surface-level diagnostics—identifying grammatical errors, punctuation flaws, and lexical redundancies—ChatGPT operates as an interactive, dialogic collaborator capable of brainstorming, outlining, restructuring arguments, and explaining rhetorical choices (Elkot et al., 2025; Hago Elmahdi et al., 2025). Understanding how these distinct technological affordances shape student writing development is essential for modernizing EFL writing curricula in Saudi Arabia.

1.2 Statement of the Problem

Despite the widespread, organic adoption of ChatGPT and Grammarly among Saudi university students, a critical pedagogical and empirical gap persists regarding their comparative effectiveness in developing specific academic writing competencies. In practice, many Saudi EFL undergraduates struggle with both macro-level writing challenges (such as thesis development, paragraph unity, and logical cohesion) and micro-level linguistic accuracy (such as syntactic complexity, tense consistency, and academic register).

While educators increasingly debate the risks of student over-reliance, uncritical copy-pasting, and academic integrity violations associated with generative AI (Hago Elmahdi et al., 2025), empirical evidence systematically comparing the structured pedagogical use of generative AI (ChatGPT) against corrective AI (Grammarly) within Saudi higher education remains scarce. Specifically, it is unclear whether generative scaffolding yields superior gains in macro-level organization and coherence compared to corrective tools, or how learners perceive the cognitive and ethical trade-offs of each platform. Without rigorous mixed-methods evidence, university faculty and curriculum designers lack clear empirical benchmarks to integrate AI tools effectively and ethically into EFL writing instruction.

1.3 Objectives of the Study

To address the identified research gap, this study pursues four primary objectives:

1. **Measure Overall Writing Proficiency Gains:** To empirically determine the extent to which the structured integration of ChatGPT and Grammarly improves overall academic writing proficiency among Saudi university EFL students compared to traditional instructor-only feedback.
2. **Compare Differential Sub-Skill Effects:** To identify statistically significant differences between ChatGPT and Grammarly across core analytical writing sub-skills, specifically *content and organization*, *coherence and cohesion*, *lexical resource*, and *grammatical accuracy and mechanics*.
3. **Examine Student Perceptions and Acceptance:** To investigate Saudi EFL learners' attitudes, perceived usefulness, and self-regulated learning experiences when utilizing generative versus corrective AI writing assistants.
4. **Formulate Pedagogical and Ethical Guidelines:** To establish evidence-based instructional frameworks and academic integrity protocols for integrating AI writing tools into Saudi tertiary EFL curricula.

1.4 Research Questions

Guided by the study's objectives, the investigation seeks to answer the following quantitative and qualitative research questions:

- **RQ1 (Overall Efficacy):** To what extent does the structured pedagogical integration of ChatGPT and Grammarly enhance overall academic writing proficiency among Saudi university EFL learners compared to conventional writing instruction?
- **RQ2 (Sub-Skill Differentiation):** Are there statistically significant differences in academic writing sub-skills (content/organization, coherence/cohesion, lexical resource, and grammatical accuracy) between Saudi EFL students trained with ChatGPT and those trained with Grammarly?
- **RQ3 (Learner Perceptions and Ethical Dimensions):** How do Saudi university EFL students perceive the pedagogical benefits, cognitive scaffolding, limitations, and ethical implications of relying on generative AI (ChatGPT) versus corrective AI (Grammarly) during the academic writing process?

1.5 Significance of the Study

This investigation offers substantial theoretical, pedagogical, and institutional contributions to applied linguistics and educational technology in the Saudi higher education context:

- **Theoretical Significance:** The study enriches second language acquisition (SLA) and computer-assisted language learning (CALL) literature by synthesizing **Vygotsky's (1978) Sociocultural Theory (Zone of Proximal Development)** with **Davis's (1989) Technology Acceptance Model (TAM)**. By conceptualizing generative and corrective AI tools as distinct forms of "digital scaffolding," the research clarifies how cognitive mediation and technological acceptance interact to shape L2 writing development.
- **Pedagogical Significance:** For EFL instructors and curriculum developers, the findings provide granular, sub-skill-level empirical evidence on *when* and *how* to deploy ChatGPT versus Grammarly across different stages of the writing process (e.g., pre-writing and structural outlining versus final proofreading and editing).
- **Institutional and Policy Significance:** In alignment with **Saudi Vision 2030** digital transformation initiatives in higher education, the study equips Saudi university administrators and quality assurance units with practical, culturally responsive guidelines to harness AI-driven personalized learning while safeguarding academic integrity and mitigating plagiarism risks (Hago Elmahdi et al., 2025).

2. Literature Review

2.1 Theoretical Framework: Scaffolding and Technology Acceptance

This research is grounded in the synthesis of two foundational frameworks:

1. **Sociocultural Theory and the Zone of Proximal Development (ZPD):** Vygotsky's (1978) framework conceptualizes learning as a socially mediated process where tools act as cultural and cognitive scaffolding. In modern digital contexts, AI writing assistants function as "digital scaffolding," bridging the gap between a learner's independent writing capability and their potential developmental level under guided support.
2. **Technology Acceptance Model (TAM):** Originally formulated by Davis (1989), TAM provides the theoretical lens to evaluate how students' *perceived usefulness* and *perceived ease of use* influence their behavioral intentions and engagement when adopting AI writing tools for academic tasks.

2.2 Generative AI (ChatGPT) in Second Language Writing

Generative pretrained transformers have transformed writing pedagogy by shifting AI from a passive dictionary or spell-checker to an active, interactive collaborator. Recent studies indicate that generative conversational AI fosters autonomy by assisting learners with idea generation, outlining, and paragraph structuring (Elkot et al., 2025; Elmahdi et al., 2025). In EFL environments, ChatGPT provides real-time, context-aware explanations that help learners understand the why behind structural revisions, thereby promoting deeper cognitive processing compared to static reference materials.

2.3 Corrective AI (Grammarly) and Automated Feedback

Automated Corrective Feedback (ACF) systems like Grammarly specialize in real-time error identification and stylistic enhancement. Extensive research demonstrates that ACF significantly reduces local errors (such as punctuation, syntax, and surface-level grammar) and encourages autonomous self-editing (Hyland & Hyland, 2019). However, scholars note a potential limitation: while corrective tools excel at micro-level accuracy, they frequently lack the contextual and rhetorical awareness required for macro-level academic coherence and argumentation, areas where generative tools or human instruction remain essential.

2.4 AI Integration in the Saudi EFL Context

As higher education in the Kingdom evolves under Saudi Vision 2030, integrating educational technology into curricula has become a national priority. Recent empirical investigations within Saudi universities highlight both the high receptivity of students toward AI-enhanced learning and the pedagogical challenges associated with over-reliance and academic integrity (Elmahdi et al., 2025). Establishing empirical benchmarks for how corrective versus generative AI tools influence specific writing competencies—such as coherence, lexical resource, and grammatical accuracy—is crucial for developing balanced, ethical institutional guidelines.

3. Research Methodology

3.1 Research Design

To rigorously evaluate the differential pedagogical impacts of generative versus corrective artificial intelligence tools on academic writing proficiency, this study employs a quasi-experimental explanatory sequential mixed-methods design (Creswell & Creswell, 2018). In the initial quantitative phase, a pre-test/post-test non-equivalent control group design was implemented across three intact academic writing sections over a 12-week semester to measure causal gains across specific writing sub-skills. In the subsequent qualitative phase, semi-structured interviews were conducted with a purposive subsample of participants to contextualize the quantitative outcomes and examine learners' cognitive scaffolding, technological acceptance, and ethical reflections.

3.2 Participants and Sampling Strategy

The target population comprised undergraduate students majoring in English Language and Translation at a public university in Saudi Arabia. Using purposive cluster sampling of three intact intermediate-to-advanced academic writing classes (N = 105 initially enrolled), participants were assigned to three instructional conditions:

- **Experimental Group 1 (EG1 – ChatGPT Group, n = 34):** Students utilized ChatGPT (GPT-4o) as a generative conversational assistant for pre-writing brainstorming, structural outlining, and macro-level revision.
- **Experimental Group 2 (EG2 – Grammarly Group, n = 33):** Students utilized Grammarly Premium as an automated corrective feedback (ACF) tool for real-time grammatical error detection, syntactic refinement, and lexical enhancement.
- **Control Group (CG – Traditional Instruction Group, n = 32):** Students received conventional instructor-led written corrective feedback (WCF) and peer-review workshops without AI tool intervention.

Following data screening for attendance compliance (minimum 85\% session attendance) and completion of both pre- and post-intervention assessments, n = 99 participants (54 females and 45 males, aged 19–23 years) were retained for final analysis. Prior to the intervention, a one-way Analysis of Variance (ANOVA) on pre-test essay scores confirmed no statistically significant baseline differences in writing proficiency across the three groups ($F(2, 96) = 0.412$, $p = 0.663$).

3.3 Data Collection Instruments

Three validated instruments were utilized to triangulate quantitative performance data with qualitative learner insights:

1. **Pre- and Post-Intervention Academic Writing Tests:** Participants completed two timed (60-minute), 350–450-word argumentative essays on comparable academic prompts before Week 1 and after Week 12. All essays were anonymized and independently scored by two experienced EFL writing instructors using a validated **20-point Analytic Writing Rubric** adapted from Jacobs et al. (1981) and the IELTS Task 2 descriptors, evaluating four dimensions (5 points each):
   - *Task Achievement, Content, and Organization*
   - *Coherence and Cohesion*
   - *Lexical Resource and Academic Register*
   - *Grammatical Range, Accuracy, and Mechanics* Inter-rater reliability was established via Intraclass Correlation Coefficient (\text{ICC} = 0.91, $p < 0.001$).
2. **AI Writing Assistant Perception and Acceptance Questionnaire:** Adapted from Davis's (1989) Technology Acceptance Model (TAM) and recent AI-in-EFL scales (Hago Elmahdi et al., 2025), this 20-item, 5-point Likert-scale instrument measured four constructs: *Perceived Usefulness* (5 items), *Perceived Ease of Use* (5 items), *Cognitive Scaffolding and Autonomy* (5 items), and *Ethical Awareness and Over-Reliance Concerns* (5 items). Internal consistency was confirmed via Cronbach's alpha (\alpha = 0.89).
3. **Semi-Structured Interview Protocol:** A 7-question interview guide was administered to 15 purposively selected students (5 from each group representing high, moderate, and low score gainers) to explore how specific AI affordances influenced their drafting, revision decisions, and critical thinking.

3.4 Instructional Intervention Procedure

The 12-week intervention was structured into three sequential phases to ensure pedagogical fidelity and prevent uncritical copy-pasting:

- **Phase 1: Orientation and Ethical Training (Week 1):** All groups completed the pre-test and received a 90-minute workshop on academic integrity, plagiarism avoidance, and responsible AI use. EG1 and EG2 received hands-on training in standardized prompting and feedback evaluation.
- **Phase 2: Guided Iterative Writing Workshops (Weeks 2–11):** Students completed four bi-weekly argumentative essay cycles following a four-stage process (*brainstorming/outlining \rightarrow first draft \rightarrow feedback & revision \rightarrow final submission*):
  - **EG1 (ChatGPT):** Students used standardized, non-generative-text prompts (e.g., *"Analyze the thesis statement and paragraph transitions in my draft and suggest structural improvements without rewriting my sentences"*) to maintain authorial ownership.
  - **EG2 (Grammarly):** Students uploaded drafts to Grammarly, reviewed automated diagnostic explanations for grammar, punctuation, and word choice, and submitted a revision log justifying which suggestions they accepted or rejected.
  - **CG (Control):** Students received marginal and end-note instructor feedback using standardized correction codes and revised their drafts manually.
- **Phase 3: Post-Testing and Qualitative Interviews (Week 12):** All participants completed the post-test essay under supervised, offline exam conditions, followed by the perception survey and individual interviews.

3.5 Data Analysis Procedures

Quantitative data were analyzed in IBM SPSS Statistics (Version 29). Descriptive statistics (means and standard deviations) were computed alongside Shapiro-Wilk normality tests. Paired-samples t-tests assessed within-group pre-to-post writing gains, while one-way Analysis of Covariance (ANCOVA)—controlling for pre-test scores—and multivariate analysis of variance (MANOVA) evaluated between-group differences across overall writing scores and individual analytical sub-skills (\alpha = 0.05, with partial eta-squared \eta_p^2 reported for effect sizes). Qualitative interview transcripts were analyzed in NVivo 14 using Braun and Clarke's (2006) six-phase thematic analysis framework.

3.6 Ethical Considerations

Ethical approval was obtained from the University's Institutional Review Board (IRB) prior to data collection. All participants signed written informed consent forms detailing voluntary participation, data anonymization, and the right to withdraw without academic penalty. To ensure equitable educational opportunity, the Control Group was granted full access to the AI writing training workshops immediately following the completion of Week 12 data collection.

4. Results and Data Analysis

This section presents the quantitative and qualitative findings derived from the 12-week quasi-experimental mixed-methods investigation (n = 99). Following the explanatory sequential framework, quantitative outcomes addressing overall writing proficiency gains (RQ1) and sub-skill differentiation (RQ2) are reported first, followed by learner perception survey diagnostics and qualitative thematic findings (RQ3).

4.1 Overall Writing Proficiency Gains (ANCOVA Results – RQ1)

Prior to conducting inferential comparisons, Shapiro-Wilk tests confirmed univariate normality across pre-test and post-test scores ($p > 0.05$), and Levene's test verified homogeneity of error variances ($F(2, 96) = 1.18$, $p = 0.312$). Furthermore, testing the interaction between the independent variable (instructional group) and the covariate (pre-test score) confirmed homogeneity of regression slopes ($F(2, 93) = 0.84$, $p = 0.435$), satisfying all assumptions for one-way Analysis of Covariance (ANCOVA).

As detailed in Table 1, paired-samples t-tests demonstrated statistically significant pre-to-post writing gains across all three groups:

- **Experimental Group 1 (EG1 – ChatGPT):** Increased from M = 11.44 (SD = 1.82) to M = 16.18 (SD = 1.65), $t(33) = 16.42$, $p < 0.001$.
- **Experimental Group 2 (EG2 – Grammarly):** Increased from M = 11.61 (SD = 1.76) to M = 15.12 (SD = 1.71), $t(32) = 12.89$, $p < 0.001$.
- **Control Group (CG – Traditional Feedback):** Increased from M = 11.38 (SD = 1.85) to M = 13.09 (SD = 1.79), $t(31) = 6.74$, $p < 0.001$.

Controlling for pre-test baseline scores, the one-way ANCOVA revealed a statistically significant main effect of instructional condition on post-test overall writing proficiency, $F(2, 95) = 38.64$, $p < 0.001$, partial \eta_p^2 = 0.449, indicating a large effect size. Bonferroni post-hoc pairwise comparisons confirmed that both EG1 (ChatGPT) (M_{\text{adj}} = 16.21, SE = 0.24, $p < 0.001$) and EG2 (Grammarly) (M_{\text{adj}} = 15.05, SE = 0.24, $p < 0.001$) significantly outperformed the Control Group (M_{\text{adj}} = 13.13, SE = 0.25). Moreover, EG1 achieved significantly higher adjusted post-test overall scores than EG2 ($p = 0.003$).

Table 1: Descriptive Statistics and One-Way ANCOVA for Overall Academic Writing Proficiency (n = 99)

Instructional Group	n	Pre-Test M (SD)	Post-Test M (SD)	Adjusted Post-Test M (SE)	Source	SS	df	MS	F	p	ηp2
EG1 (ChatGPT)	34	11.44 (1.82)	16.18 (1.65)	16.21 (0.24)	Pre-Test (Covariate)	94.12	1	94.12	48.27	< 0.001	0.337
EG2 (Grammarly)	33	11.61 (1.76)	15.12 (1.71)	15.05 (0.24)	Group Condition	150.70	2	75.35	38.64	< 0.001	0.449
CG (Control)	32	11.38 (1.85)	13.09 (1.79)	13.13 (0.25)	Error	185.25	95	1.95			

4.2 Differential Effects Across Writing Sub-Skills (MANOVA Results – RQ2)

To examine how generative versus corrective AI tools influenced distinct analytical dimensions of academic writing, a one-way Multivariate Analysis of Variance (MANOVA) was conducted across the four post-test rubric sub-skills (scored out of 5.00 each). Box's M test confirmed equality of covariance matrices ($M = 24.18$, $F(20, 32410) = 1.15$, $p = 0.289$).

The multivariate test yielded a statistically significant difference among the three instructional groups across the combined sub-skills, Wilks' \Lambda = 0.362, $F(8, 186) = 15.41$, $p < 0.001$, partial \eta_p^2 = 0.399. As presented in Table 2, univariate follow-up tests and Bonferroni post-hoc analyses uncovered a clear functional divergence between the two AI interventions:

1. **Macro-Level Sub-Skills (Content/Organization & Coherence/Cohesion): EG1 (ChatGPT)** achieved the highest mean scores in *Content and Organization* (M = 4.18, SD = 0.52) and *Coherence and Cohesion* (M = 4.12, SD = 0.49), significantly outperforming both **EG2 (Grammarly)** (M = 3.58 and M = 3.52, respectively; $p < 0.001$) and the **Control Group** (M = 3.25 and M = 3.19; $p < 0.001$).
2. **Micro-Level Sub-Skills (Grammatical Accuracy & Mechanics): EG2 (Grammarly)** recorded the highest mean score in *Grammatical Range, Accuracy, and Mechanics* (M = 4.24, SD = 0.46), slightly surpassing **EG1 (ChatGPT)** (M = 3.94, SD = 0.51, $p = 0.042$) and substantially outperforming the **Control Group** (M = 3.31, SD = 0.54, $p < 0.001$).
3. **Lexical Resource:** Both **EG1** (M = 3.94, SD = 0.48) and **EG2** (M = 3.78, SD = 0.50) significantly exceeded the **Control Group** (M = 3.34, SD = 0.52, $p < 0.001$), with no statistically significant difference between the two AI groups ($p = 0.548$).

Table 2: MANOVA Univariate Comparisons Across Analytical Writing Sub-Skills (n = 99)

Writing Sub-Skill (Max = 5.00)	EG1: ChatGPT (n=34) M (SD)	EG2: Grammarly (n=33) M (SD)	CG: Control (n=32) M (SD)	F(2,96)	p-Value	ηp2	Post-Hoc Group Differences (Bonferroni)
Content & Organization	4.18 (0.52)	3.58 (0.55)	3.25 (0.57)	24.85	< 0.001	0.341	\text{EG1} > \text{EG2}^* > \text{CG}^*
Coherence & Cohesion	4.12 (0.49)	3.52 (0.53)	3.19 (0.56)	26.42	< 0.001	0.355	\text{EG1} > \text{EG2}^* > \text{CG}^*
Lexical Resource & Register	3.94 (0.48)	3.78 (0.50)	3.34 (0.52)	12.64	< 0.001	0.208	\text{EG1}, \text{EG2} > \text{CG}^* (\text{EG1} \approx \text{EG2})
Grammatical Accuracy & Mechanics	3.94 (0.51)	4.24 (0.46)	3.31 (0.54)	29.18	< 0.001	0.378	\text{EG2} > \text{EG1}^* > \text{CG}^*

Note. * denotes statistically significant pairwise differences at $p < 0.05$.

4.3 Student Perceptions and Qualitative Thematic Findings (RQ3)

4.3.1 Quantitative Survey Diagnostics

Descriptive analysis of the post-intervention TAM perception questionnaire among experimental participants (n = 67) indicated high overall acceptance across both platforms, alongside distinct functional preferences (Table 3). Students in EG1 rated ChatGPT higher for Cognitive Scaffolding and Autonomy (M = 4.38, SD = 0.54) than EG2 rated Grammarly (M = 3.82, SD = 0.61; t(65) = 3.98, p < 0.001), whereas EG2 reported higher Perceived Ease of Use (M = 4.52, SD = 0.45) and lower Ethical and Over-Reliance Anxiety (M = 2.41, SD = 0.72) compared to EG1 (M = 3.64, SD = 0.78; t(65) = 6.71, p < 0.001).

Table 3: Experimental Group Perceptions on TAM and Ethical Dimensions (n = 67)

Perception Construct (5-Point Scale)	EG1: ChatGPT (n=34) M (SD)	EG2: Grammarly (n=33) M (SD)	t(65)	p-Value
Perceived Usefulness	4.41 (0.51)	4.29 (0.53)	0.94	0.351
Perceived Ease of Use	4.15 (0.58)	4.52 (0.45)	-2.91	0.005
Cognitive Scaffolding & Autonomy	4.38 (0.54)	3.82 (0.61)	3.98	< 0.001
Ethical & Over-Reliance Concerns	3.64 (0.78)	2.41 (0.72)	6.71	< 0.001

4.3.2 Qualitative Thematic Analysis

Thematic analysis of the 15 semi-structured interview transcripts in NVivo 14 yielded four overarching themes that explain the quantitative divergence between generative and corrective AI scaffolding:

- **Theme 1: Macro-Structural Dialogic Scaffolding vs. Micro-Surface Polishing.** Participants in EG1 consistently highlighted ChatGPT's capacity to act as a conversational brainstorming and structural tutor, helping them organize thesis arguments and paragraph transitions before and during drafting. Conversely, EG2 participants described Grammarly as a precision proofreading assistant that caught subtle article, preposition, and subject-verb agreement errors.

"When I used ChatGPT with the structural prompt, it showed me why my second body paragraph did not connect to my thesis, without writing the sentences for me." (Participant 4, EG1)

"Grammarly immediately highlighted my tense shifts and repetitive words while I typed, which made my final sentences much cleaner." (Participant 9, EG2)

- **Theme 2: Metalinguistic Awareness and Self-Regulated Revision.** Students across both experimental conditions reported that instant, explanatory feedback accelerated their metalinguistic awareness compared to waiting weeks for

instructor comments. In EG2, maintaining a revision log forced students to evaluate whether Grammarly's stylistic suggestions preserved their intended academic meaning.

"At first I clicked 'accept' on every Grammarly suggestion, but keeping the revision log made me realize some suggestions changed my academic tone, so I learned to decide for myself." (Participant 11, EG2)

- **Theme 3: Ethical Vigilance and Authorial Ownership.** A recurring concern among EG1 learners was navigating the boundary between legitimate structural guidance and uncritical text generation. Students emphasized that structured prompt training was essential to prevent dependency and uphold academic integrity.

"Without the instructor's prompt rules, it is very tempting to let ChatGPT write whole paragraphs. Using diagnostic prompts kept the ideas mine." (Participant 2, EG1)

- **Theme 4: Desire for Sequenced Hybrid Integration.** Notably, participants from all three groups—including Control Group students interviewed after their post-intervention workshop—advocated for a sequenced pedagogical workflow combining both tools: utilizing ChatGPT during the pre-writing and structural outlining stages, followed by Grammarly during final linguistic editing.

"If we could use ChatGPT first to organize our ideas and arguments, and then run the final draft through Grammarly for grammar, our academic essays would be much stronger." (Participant 7, EG1)

5. Discussion, Pedagogical Implications, and Conclusion

5.1 Discussion of Findings

This mixed-methods investigation examined the differential impacts of generative AI (ChatGPT) and automated corrective AI (Grammarly) on developing academic writing proficiency among Saudi university EFL students. The quantitative ANCOVA and MANOVA outcomes, triangulated with qualitative interview themes, demonstrate that while both AI platforms substantially outperform traditional instructor-only feedback, they operate through distinct cognitive and pedagogical mechanisms.

5.1.1 Overall Writing Gains and Digital Scaffolding (RQ1)

The one-way ANCOVA confirmed that both EG1 (ChatGPT, M_{\text{adj}} = 16.21) and EG2 (Grammarly, M_{\text{adj}} = 15.05) achieved significantly greater overall post-test writing gains than the Control Group (M_{\text{adj}} = 13.13, \eta_p^2 = 0.449). Viewed through Vygotsky's (1978) Sociocultural Theory and Zone of Proximal Development (ZPD), these results indicate that immediate, iterative AI feedback functions as an effective "digital scaffold" that bridges the gap between Saudi EFL learners' independent writing capabilities and higher academic proficiency. Unlike delayed traditional written corrective feedback—which often arrives after students have moved on from the cognitive drafting process (Hyland & Hyland, 2019)—real-time AI interaction enables learners to test hypotheses, revise arguments, and internalize rhetorical conventions while actively engaged in composition (Elkot et al., 2025). Notably, ChatGPT yielded superior overall adjusted gains compared to Grammarly ($p = 0.003$), largely because macro-structural improvements carry foundational weight across holistic academic writing quality.

5.1.2 Functional Divergence Across Analytical Writing Sub-Skills (RQ2)

The MANOVA results (\text{Wilks' } \Lambda = 0.362, \eta_p^2 = 0.399) revealed a clear division of pedagogical labor between generative and corrective AI tools:

- **Macro-Level Rhetorical Gains via Generative AI:** EG1 (ChatGPT) significantly outperformed EG2 (Grammarly) in *Content and Organization* ($M = 4.18$ vs. 3.58) and *Coherence and Cohesion* ($M = 4.12$ vs. 3.52). As corroborated by **Theme 1 (Macro-Structural Dialogic Scaffolding vs. Micro-Surface Polishing)**, ChatGPT's conversational architecture allowed learners to interrogate thesis alignment, paragraph transitions, and logical progression. Within Vygotsky's ZPD framework, dialogic prompting simulated peer-or-tutor negotiation, prompting students to critically evaluate macro-level text structure rather than focusing solely on sentence-level mechanics.
- **Micro-Level Linguistic Precision via Corrective AI:** Conversely, EG2 (Grammarly) achieved the highest proficiency in *Grammatical Range, Accuracy, and Mechanics* ($M = 4.24$), significantly surpassing EG1 ($M = 3.94$, $p = 0.042$). Consistent with automated corrective feedback literature (Hyland & Hyland, 2019), Grammarly's rule-based and neural diagnostic engine excelled at identifying local syntactic errors, tense inconsistencies, and punctuation flaws. Furthermore, **Theme 2 (Metalinguistic Awareness and Self-Regulated Revision)** demonstrates that pairing Grammarly with a mandatory revision log prevented passive click-acceptance and fostered critical metalinguistic reflection.
- **Convergent Lexical Enhancement:** Both EG1 ($M = 3.94$) and EG2 ($M = 3.78$) produced comparable, statistically significant gains in *Lexical Resource* over the Control Group ($M = 3.34$), showing that both contextual synonyms (Grammarly) and dialogic vocabulary suggestions (ChatGPT) effectively expand Saudi learners' academic register (Alshammari et al., 2022).

5.1.3 Technology Acceptance, Ethical Vigilance, and Hybrid Synergy (RQ3)

Interpreted through Davis's (1989) Technology Acceptance Model (TAM), survey and interview data explain why students engaged differently with each platform. While both groups reported comparably high Perceived Usefulness ($M = 4.41$ for ChatGPT; $M = 4.29$ for Grammarly), students rated Grammarly significantly higher in Perceived Ease of Use ($M = 4.52$ vs. 4.15, $p = 0.005$) due to

its non-intrusive, inline interface. In contrast, ChatGPT required deliberate prompt engineering, which—when structured via non-generative diagnostic prompts—yielded significantly higher Cognitive Scaffolding and Autonomy ($M = 4.38$ vs. 3.82, $p < 0.001$).

Simultaneously, Theme 3 (Ethical Vigilance and Authorial Ownership) and elevated Ethical and Over-Reliance Concerns in EG1 ($M = 3.64$ vs. 2.41) highlight a critical pedagogical tension: generative AI offers deeper cognitive scaffolding but carries higher risks of uncritical text generation if left unguided (Hago Elmahdi et al., 2025). Crucially, Theme 4 (Desire for Sequenced Hybrid Integration) points toward an optimal pedagogical resolution—deploying ChatGPT during early ideation and structural outlining, followed by Grammarly during final linguistic editing.

5.2 Pedagogical and Institutional Implications

The findings offer actionable recommendations for EFL faculty, curriculum designers, and higher education policymakers in Saudi Arabia:

1. **Adopt a Sequenced Hybrid AI Writing Workflow:** Rather than treating ChatGPT and Grammarly as interchangeable or mutually exclusive, writing instructors should integrate both tools sequentially across the process-writing cycle:
   - *Stage 1 (Pre-Writing & Macro-Revision):* Utilize **ChatGPT** with standardized, non-text-generating diagnostic prompts to evaluate thesis clarity, argument logic, and inter-paragraph cohesion.
   - *Stage 2 (Micro-Editing & Proofreading):* Utilize **Grammarly** to refine grammatical accuracy, syntactic variety, and academic word choice.
2. **Mandate Prompt Literacy and Reflective Revision Logs:** To safeguard academic integrity and prevent passive dependency, universities should incorporate explicit training in ethical prompt engineering alongside reflective revision logs, requiring students to justify accepted or rejected AI suggestions.
3. **Update Institutional AI Assessment Policies under Saudi Vision 2030:** Academic departments should transition from blanket bans on AI tools to transparent, rubric-aligned AI-assisted writing policies that evaluate both the final written product and the documented revision process, preparing Saudi graduates for AI-augmented professional environments.

5.3 Limitations and Directions for Future Research

Several limitations warrant consideration when interpreting these findings. First, the quasi-experimental intervention spanned a single 12-week semester at one Saudi public university; longitudinal studies tracking students across multiple academic years are needed to verify long-term retention of writing gains when AI scaffolds are removed. Second, this study isolated ChatGPT and Grammarly into separate experimental conditions to measure their distinct sub-skill effects; future experimental research should directly test a combined hybrid group (ChatGPT + Grammarly) against single-tool cohorts. Finally, incorporating keystroke logging or screen-recording analytics would provide granular behavioral evidence of how learners interact with AI feedback in real time.

5.4 Conclusion

As generative and corrective artificial intelligence tools reshape higher education, effective EFL writing pedagogy depends on aligning specific technological affordances with targeted linguistic and rhetorical goals. Drawing on mixed-methods evidence from 99 Saudi university EFL undergraduates, this study demonstrates that ChatGPT serves as a powerful dialogic scaffold for macro-level content organization and textual cohesion, whereas Grammarly excels as a precision instrument for micro-level grammatical accuracy and stylistic polish. When anchored in ethical prompt literacy, metalinguistic reflection, and clear institutional governance, the strategic integration of AI writing assistants can substantially elevate academic writing proficiency while preserving student agency and scholarly integrity.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Alshammari, A., Alenezi, M., & Almutairi, H. (2022). Challenges in Saudi EFL academic writing: Structural cohesion, syntactic accuracy, and pedagogical interventions. *Arab World English Journal*, *13*(2), 45–60.

[2] Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

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

[4] 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

[5] Elkot, M. A., Youssif, E., Hago Elmahdi, O. E., AbdAlgane, M., & Ali, R. (2025). Generative conversational AI: Active practices for fostering students with mild intellectual disabilities to improve English communication skills. *Contemporary Educational Technology*, *17*(1), ep549. https://doi.org/10.30935/cedtech/15688

[6] Hago Elmahdi, O. E., Allehyani, F. S., & AbdAlgane, M. (2025). Exploring the ethical implications and cultural considerations in using AI technologies for EFL education in the Saudi context. *Nsukka Journal of the Humanities*, *33*(2), 1–18.

[7] Hago Elmahdi, O. E., Allehyani, F. S., & AbdAlgane, M. (2025). Exploring the ethical implications and cultural considerations in using AI technologies for EFL education in the Saudi context. *Nsukka Journal of the Humanities*, *33*(2), 1–18.

[8] Hyland, K., & Hyland, F. (Eds.). (2019). *Feedback in second language writing: Contexts and issues* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/9781108635547

[9] Jacobs, H. L., Zinkgraf, S. A., Wormuth, D. R., Hartfiel, V. F., & Hughey, J. B. (1981). Testing ESL composition: A practical approach. Newbury House.

[10] Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes (M. Cole, V. John-Steiner, S. Scribner, & E. Souberman, Eds.)*. Harvard University Press.

Appendix A: 20-Point Analytic Writing Rubric

Criterion (5 Points Each)	5 – Exemplary	4 – Proficient	3 – Developing	2 – Emerging	1 – Inadequate
1. Task Achievement, Content & Organization	Insightful, fully developed thesis; thoroughly substantiated arguments; clear introduction, logical body progression, and compelling conclusion.	Clear thesis and relevant arguments with solid supporting details; coherent overall essay structure with minor lapses in depth.	Identifiable thesis; arguments are present but occasionally underdeveloped or repetitive; basic organizational structure.	Unclear or overly broad thesis; insufficient supporting evidence; weak paragraph focus or abrupt structural shifts.	Fails to address the prompt; lacks a discernible thesis; disorganized ideas with minimal supporting detail.
2. Coherence & Cohesion	Seamless logical flow; sophisticated and varied cohesive devices; strong inter- and intra-paragraph transitions; clear referencing.	Clear logical progression; effective use of cohesive devices, though occasionally mechanical or repetitive; clear paragraphing.	Generally coherent, but transitions between ideas or paragraphs are sometimes abrupt or over-reliant on basic connectors.	Frequent breaks in logical flow; inaccurate or missing cohesive devices; unclear pronoun or lexical referencing.	Disjointed sentences lacking logical progression; absence of effective cohesive links; difficult to follow.
3. Lexical Resource & Academic Register	Wide, sophisticated range of academic vocabulary and collocations; precise word choice; consistent formal scholarly register.	Sufficient academic vocabulary allowing flexibility and precision; rare word-choice errors that do not impede meaning; appropriate register.	Adequate general and academic vocabulary; occasional awkward phrasing, repetition, or minor register shifts.	Limited vocabulary range; frequent errors in word form, collocation, or register that occasionally obscure meaning.	Very restricted vocabulary; pervasive word-choice errors and informal expressions severely impeding comprehension.
4. Grammatical Range, Accuracy & Mechanics	Wide variety of complex syntactic structures; consistently accurate grammar, punctuation, and spelling; rare minor slips.	Good control of complex and simple structures; frequent error-free sentences; minor grammatical or punctuation errors.	Mix of simple and complex sentences; noticeable grammatical, tense, or punctuation errors that rarely obscure meaning.	Predominantly simple structures; frequent errors in verb tense, agreement, articles, and punctuation causing strain.	Severe, pervasive grammatical and mechanical errors distorting meaning throughout the essay.

Appendix B: Standardized Prompts & Reflective Revision Log

Part 1: Standardized Non-Generative ChatGPT Prompts (EG1)

To uphold academic integrity and preserve authorial ownership, EG1 participants were restricted to the following diagnostic prompts:

Pre-Writing & Outlining Prompt:

"I am writing a 400-word academic argumentative essay on [Topic]. My proposed thesis statement is: '[Insert Thesis]'. Evaluate the clarity, arguability, and academic focus of this thesis, and suggest three logical sub-arguments I could develop into body paragraphs. Do not write the essay or full paragraphs for me."

Macro-Structural & Cohesion Diagnostic Prompt:

"Act as an academic writing tutor. Analyze the draft below strictly for paragraph unity, logical progression, and transitions between paragraphs. Identify where my argument lacks cohesion or supporting evidence, and ask me guiding questions to help me revise it myself. Do not rewrite my sentences: [Insert Draft]."

Academic Register & Lexical Feedback Prompt:

"Identify up to five informal words or repetitive phrases in my draft below, explain why they are less suitable for academic writing, and suggest three formal academic alternatives for each so I can choose the best fit: [Insert Draft]."

Part 2: Grammarly Reflective Revision Log Template (EG2)

Draft Section / Sentence	**Original Text**	**Grammarly Diagnostic Category (Grammar / Clarity / Engagement / Delivery)**	**Suggested Revision**	**Decision (Accepted / Modified / Rejected)**	**Pedagogical Justification (Why did you accept, modify, or reject this change?)**
Example: Body Para 1, Sent. 2	Students was affected by delay feedback.	Grammar (Subject-Verb Agreement & Word Form)	Students were affected by delayed feedback.	Accepted	Corrected plural verb agreement ('were') and used adjective form ('delayed').
Example: Body Para 2, Sent. 4	Furthermore, this paradigm proves vital.	Clarity (Word Choice)	Also, this model is important.	Rejected	Grammarly's suggestion lowered the formal academic register of my argument.
1.					
2.					
3.					