Article

Check for updates

Journal of International Students
Volume 16, Issue 15 (2026), pp. 81-102
ISSN: 2162-3104 (Print), 2166-3750 (Online)
jistudents.org
https://doi.org/10.32674/e7jdhc67

Structured AI Workflows in Saudi EMI Research Writing: Performance Gains, Equity, and Academic Integrity Across Three Semesters

Erick C.W. Nelson
Prince Sultan University, Saudi Arabia
ORCID: 0009-0002-4011-5120

ABSTRACT: *Despite growing interest in generative AI for writing instruction, empirical evidence from English-medium programs in Arabic-speaking Gulf contexts remains scarce. This mixed-methods case study examines a staged integration of generative AI (GenAI) into a core English course, Research Writing Techniques at a private English-medium university in Saudi Arabia (English-Medium Instruction; EMI) that enrolls predominantly first-year Arabic L1 students. Across three consecutive semesters, the intervention progressed from AI-enhanced instructor materials (Semester 241), to optional student-facing support via a course-specific CustomGPT tutor and short recap videos (Semester 242), to full Week-1 integration of a mandatory homework pipeline in which students consulted the course Guide, practiced with the CustomGPT tutor, reviewed via Quizlet, and completed proctored quizzes through Examplify, alongside a curricular shift in Assignment 1 from an Annotated Bibliography to a Literature Review (Semester 251). Data sources included ExamSoft midterm and final exam scores, quiz gradebooks, aggregate CustomGPT usage metrics, YouTube analytics, student surveys, and brief email interviews. Compared to baseline, Semester 251 students scored higher on homework—even under stricter, proctored conditions (76.00%* vs. *94.61%; Hedges 'g = 1.20) and moderate gains on comprehensive final exams (61.36%* vs. *70.46%; g = 0.52). A large midterm effect (64.29%* vs. *85.01%; g = 1.19) and reduced score variability indicate that the structured workflow promoted deep, more equitable learning in APA citation and research methods. Survey and interview data show that students primarily used GenAI for brainstorming, outlining, and revision, endorsed an "AI as coach, not ghostwriter" stance, and valued optional Arabic explanations for technical concepts. The results show that a coherent package of CustomGPT tutoring and brief recap videos can enhance first-year research-writing outcomes and*

academic integrity practices in English-medium settings.

Keywords: academic integrity, CustomGPT, English-medium instruction, generative AI, research writing

Received: Dec 12, 2025 | **Revised**: April 13, 2026 | **Accepted:** April 18, 2026

How to Cite: 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

INTRODUCTION

Generative AI (GenAI) has quickly become an integral part of writing pedagogy, offering scalable scaffolds for idea generation, planning, and revision while raising questions about authorship and assessment integrity (Freeman, 2025; Mollick & Mollick, 2024; Qian et al., 2025; Wang & Fan, 2025). This paper documents a staged, classroom-based implementation in a core English course, Research Writing Techniques at a private English-medium university in Riyadh, Saudi Arabia. The context is key: Research Writing Techniques enroll primarily Arabic L1 (first-language) first-year students studying through English-Medium Instruction (EMI). The course is taught entirely in English; however, optional Arabic explanations are available via a course-specific CustomGPT to address key challenges for multilingual learners.

While the Institute of International Education (2024) defines an international student as 'anyone studying at an institution of higher education in the United States on a temporary visa,' the scope of international education has expanded to include 'Internationalization at Home' (IaH). This study adopts a broader definition of international student experience, focusing on English-medium instruction (EMI) learners who navigate linguistic and academic border crossing within their home institutions. As global student mobility rebounds to prepandemic levels (UNESCO, 2023), understanding the digital workflows of these 'at-home' international learners is critical for equitable pedagogy.

The intervention evolved across three academic terms at the university to systematically enhance the learning environment: Semester 241 (Fall 2024), Semester 242 (Spring 2025), and Semester 251 (Fall 2025). In Semester 241, the instructor used AI only to refine and streamline instructor-facing lecture materials.

Semester 242 introduced student-facing AI support, making a CustomGPT tutor and concise Synthesia recap videos available. The system culminated in Semester 251 with full Week-1 integration, implementing a mandatory, structured homework pipeline in which students first consulted the course Guide, then practiced with the CustomGPT tutor, reviewed terminology via Quizlet, and completed a 10-item, single-attempt proctored quiz through Examplify. This final phase also marked a key curricular change, shifting the first major assignment (AS1) from an annotated bibliography to a literature review.

Research writing techniques are designed to build foundational knowledge in core research concepts (e.g., quantitative vs. qualitative designs, sampling), APA style, academic integrity, and scholarly communication. The course assessment arc comprises three major assignments: the Literature Review (AS1), followed by the Research Essay (AS2), and culminating in the Research Poster and Exhibition (AS3). Given the phased integration of integrity-by-design AI supports within this EMI environment, the objective of this study is to examine how this coherent package of tools and structured workflows relates to student performance and learner perceptions over the tri-semester period.

Specifically, this study addresses the following research questions:

RQ_1: How do successive Research Writing Techniques cohorts perceive the usefulness of CustomGPT, recap videos, and AI-enhanced materials?

RQ_2: To what extent does the structured homework workflow (Guide, CustomGPT, Quizlet, and Examplify, single attempt) relate to quiz and exam performance in Fall 2025?

RQ_3: How do students report handling academic integrity concerns when using AI (e.g., coaching vs. copying; optional Arabic explanations)?

LITERATURE REVIEW

Generative AI has rapidly expanded the toolkit available to writing instructors, but empirical evidence from English-medium programs in the Arab region remains limited. Sector-wide surveys indicate that students most often use GenAI to help explain difficult concepts, summarize readings, generate research ideas, and edit or refine drafts rather than to produce entire essays end-to-end (Freeman, 2025). A survey of 245 international students in the U.S. found that while perceived usefulness, ease of use, and enjoyment positively predicted GenAI adoption, fear of plagiarism had a significant negative effect on students' intention to use GenAI, a concern particularly pronounced among nonnative English speakers (Ittefaq et al., 2025). This suggests that clear, structured workflows are essential to reduce anxiety and encourage ethical engagement with these tools. A systematic review by Luo et al. (2025) highlights that the effective implementation of AI-based tools in higher education hinges on design strategies that offer benefits such as personalized learning experiences and real-time feedback. At the same time, scholarship on English-medium instruction (EMI) in

Saudi Arabia highlights persistent challenges for multilingual learners in academic writing (particularly argumentation, citation practice, and genre awareness), underscoring the need for explicit, scaffolded support (Elyas & Al-Hoorie, 2024; Macaro et al., 2018). Recent scholarship in the *Journal of International Students* further argues that AI tools hold particular promise for advancing equity and inclusion among international student populations, especially when designed to address language barriers through L1 support and personalized feedback mechanisms (Duke, 2026).

Two complementary theoretical lenses motivate this study's design. First, the Technology Acceptance Model (TAM) predicts that students 'intention to adopt an instructional technology is driven by perceived usefulness and perceived ease of use (Davis, 1989). Aligning GenAI-supported workflows with the actual demands of Research Writing Techniques tasks (e.g., generating search strings, building outline skeletons tied to sources, paraphrasing with anchor citations, or drafting topic-sentence frames) can therefore elevate perceived usefulness and sustain adoption. Second, cognitive load theory and multimedia learning suggest that well-timed, segmented explanations reduce extraneous processing and free capacity for schema construction (Sweller, 1988; Mayer, 2009). Short recap videos and microprompts that target a single subskill (e.g., turning claims into APA-cited sentences) are thus expected to be more effective than longer, undifferentiated demonstrations. Evidence from large-scale massive open online course (MOOC) data similarly shows that concise, focused videos are associated with higher engagement than longer lectures (Guo et al., 2014), while experimental evidence confirms that microlearning integration in digital platforms positively affects student engagement in higher education (Muali & Karlina, 2025).

Emerging syntheses of GenAI in education further situate the present work. Meta-analytic evidence suggests that access to ChatGPT can produce positive, small-to-moderate effects on learning performance, perceptions, and higher-order thinking, with substantial heterogeneity by task and scaffolding (Wang & Fan, 2025). Systematic reviews focused on language education in 2023–2024 report a rapid shift from proof-of-concept to classroom-embedded studies but also highlight design gaps, especially the need for transparent prompts, process-based assessments, and clear academic integrity guidance (Qian et al., 2025; Mogavi et al., 2024; Mollick & Mollick, 2024). To align with this evidence, the first phase of the present tri-semester design used GenAI to enhance and structure existing instructor PPTs (clarifying task sequences, exemplars, and checks for understanding) before deploying CustomGPT tutors and brief recap videos to reinforce the refined course architecture.

A parallel line of evidence cautions against overreliance on AI-text detection. Peer-reviewed studies document that popular detectors can exhibit nontrivial false-positive rates, with disproportionately higher misclassification risk for nonnative English writers (Liang et al., 2023). In line with this evidence, the study emphasizes process evidence (drafts, source notes, and work logs) and assignment designs that require citation-linked reasoning over policing model outputs.

These studies establish four key findings: (a) adoption hinges on perceived usefulness and ease within authentic writing tasks (Davis, 1989), (b) segmented, low-load microsupports improve engagement and learning (Mayer, 2009; Guo et al., 2014; Muali & Karlina, 2025), (c) multilingual writers in Saudi EMI contexts benefit from explicit scaffolds (Elyas & Al-Hoorie, 2024; Macaro et al., 2018), and (d) detector limitations necessitate a process-evidence paradigm (Liang et al., 2023).

METHODS

This study used a mixed-methods, quasiexperimental cohort design to examine the staged integration of generative AI into a core English course, Research Writing Techniques at a private English-medium university in Saudi Arabia. Three consecutive cohorts (Semesters 241, 242, and 251) were taught by the same instructor under progressively more intensive AI conditions. Semester 241 served as a baseline, with GenAI used primarily to enhance instructor-facing materials. Semester 242 added optional student-facing support through a course-specific CustomGPT tutor and brief recap videos. Semester 251 implemented a structured AI-supported homework workflow in which students first consulted the course Guide, then practiced with the CustomGPT tutor, reviewed terminology via Quizlet, and completed a proctored quiz through Examplify, alongside a redesigned first major assignment focused on the literature review. The following subsections describe the participants, instructional context, instruments, data sources, and analytic procedures used to address the three research questions.

Participants

Three independent cohorts of first-year undergraduates enrolled in Research Writing Techniques participated in this study: Semester 241 (Fall 2024, N = 67), Semester 242 (Spring 2025, N = 82), and Semester 251 (Fall 2025, N = 82), across sections 796, 799, 800, and 801. All students studied through English-medium instruction, with Arabic as their primary home language.

Table 1: ***Sample Characteristics and Instructional Conditions Across Three Semesters***

Semester	Term	N	Intervention Phase	Quiz Structure
241	Fall 2024	67	Instructor-only AI materials enhancement (baseline)	5-item, 3-attempt, unproctored LMS
242	Spring 2025	82	CustomGPT + recap videos introduced; same quiz	5-item, 3-attempt, unproctored LMS

Semester	Term	N	Intervention Phase	Quiz Structure
251	Fall 2025	82	Full integration: mandatory Guide→CustomGPT→ Quizlet→ Examplify workflow; AS1 shift to Literature Review	10-item, 1-attempt, proctored Examplify

Note. LMS = Learning Management System; AS1 = Assignment 1. All participants were first-year undergraduates enrolled in Research Writing Techniques at a private Saudi university. The increase in quiz difficulty from Semester 242 to 251 (doubling items, eliminating retakes, requiring proctoring) represents substantially higher stakes despite continued AI tool availability. Semester 241 had one student absence on the final exam (N = 66 of 67 analyzed). Semester 242 had one student who received DN (Denial) status prior to the final exam period but was permitted to take the exam; this student was excluded from all analyses due to the absence of coursework data (N = 82 analyzed for all assessments).

Instructional Conditions

Semester 241. Instructor-only AI was used for materials enhancement, and students completed five-item, three-attempt quizzes in the learning management system (LMS).

Semester 242. The same core assessments were retained, while a course-specific CustomGPT and recap videos were made available; informal in-class feedback was gathered.

Semester 251. AI-enhanced slides, a course CustomGPT, a recap set, and a homework pipeline in which students consulted the course Guide, practiced with the CustomGPT tutor, reviewed via Quizlet, and completed a single-attempt quiz through Examplify were implemented; Assignment 1 shifted to a Literature Review with NotebookLM supporting source comprehension.

Data Sources

Quantitative Data Sources

Assessments. ExamSoft midterm and final examinations were used across semesters (Semester 241: final exam only, $N = 66$; Semester 242: midterm and final, $N = 82$; Semester 251: midterm and final, $N = 82$). Final exams assessed identical course content but differed in format: Semesters 241--242 included 17 items (7 short-answer, 10 multiple-choice), whereas Semester 251 included 9

items (4 essay, 5 multiple-choice); all exams totaled 40 points and were administered via proctored ExamSoft.

Usage indicators. YouTube playlist metrics were collected for the designated recap series (shorts excluded, legacy annotated-bibliography videos treated as extraneous), and aggregate CustomGPT session counts were tracked (no per-student logs).

Surveys. A Semester 242 perceptions survey covered CustomGPT, recaps, and materials; Semester 251 surveys (main plus add-on) covered workflow, Prompt Assists/Lab, NotebookLM, and Arabic usage with CustomGPT.

Qualitative Data Sources

Interviews Short email interviews were conducted with students from both Semester 242 and Semester 251 (e.g., S-242-Email-01).

Artifacts. Guides, rubrics, templates, work-logs, Prompt Lab materials, and lecture decks were archived as instructional artifacts.

Procedure

Quantitative Data Collection.

Assessment data were collected automatically through ExamSoft at the end of each semester. Quiz scores were exported from the LMS (Semesters 241 and 242) and Examplify (Semester 251). YouTube analytics and aggregate CustomGPT session counts were collected at the close of each semester. Survey responses were gathered electronically via the LMS.

Qualitative Data Collection.

Email interviews were conducted voluntarily at the end of Semesters 242 and 251. Students were invited to respond to open-ended questions about their experience with the AI tools, academic integrity, and workflow. Instructional artifacts, including guides, rubrics, work logs, and lecture decks, were archived throughout the intervention period.

Ethics

This study received expedited approval from the Prince Sultan University Institutional Review Board (IRB Reference No.: PSU IRB-2025-11-0266, approved 9 November 2025). Electronic consent preceded surveys; interviews were voluntary by email. Data are deidentified and stored securely. Participation is voluntary, does not affect course grades, and carries minimal risk; participants may skip any item.

Analysis Plan

This study employed a convergent mixed-methods design, collecting quantitative performance data and qualitative perception data concurrently, with integration occurring in the Discussion through side-by-side comparison of statistical outcomes and thematic findings. Quantitative analyses were conducted using Microsoft Excel. Absences and empty submissions were excluded from central-tendency calculations, with valid N and missingness reported for each measure. Descriptive statistics (N, mean, SD) were calculated by cohort for homework, midterm, and final exam scores. Between-cohort comparisons (241→242 and 242→251) were conducted using independent-samples t tests, with mean differences, 95% confidence intervals, and Hedges' g reported; Hedges' g was chosen over Cohen's d due to unequal sample sizes. Effect sizes were interpreted using established conventions: small ($g \approx 0.2$), medium ($g \approx 0.5$), and large ($g \approx 0.8$) (Cohen, 1988). Usage indicators (CustomGPT conversation counts and YouTube analytics) were summarized descriptively. Qualitative data from email interviews were analyzed using inductive thematic coding and mapped to survey constructs, including tutor helpfulness, recap utility, integrity strategies, and Arabic explanation use.

RESULTS

Homework and Quiz Performance

The homework component evolved across the three semesters in both assessment structure and pedagogical support. Semester 241 (baseline) and Semester 242 both used 5-item, three-attempt LMS quizzes, allowing students multiple opportunities to demonstrate mastery. Semester 251 shifted to 10-item, single-attempt Examplify quizzes administered in a proctored environment, representing substantially increased difficulty and higher stakes.

Semester 241 to 242 (Introduction of AI Tools).

Semester 242 introduced the CustomGPT tutor and Synthesia recap videos while maintaining the same quiz structure as Semester 241. Homework performance improved dramatically: from $M = 76.00\%$ ($SD = 20.61$, $N = 67$) in Semester 241 to $M = 99.60\%$ ($SD = 2.47$, $N = 82$) in Semester 242, representing a gain of 23.60 percentage points. This constitutes a very large effect size (Hedges' $g = 1.70$, 95% CI [1.32, 2.08]). Notably, the standard deviation decreased from 20.61 to 2.47, indicating that AI tool availability not only raised the mean but also substantially reduced performance variability—suggesting that the tools served as an equity mechanism by helping lower-performing students reach near-ceiling performance.

Semester 242 to 251 (Structured Workflow Under Stricter Conditions).

Despite substantially increased quiz difficulty, individual quiz means in Semester 251 remained consistently high: HW1 (M = 97.32%, SD = 6.10), HW2 (M = 93.54%, SD = 14.00), HW3 (M = 93.66%, SD = 13.38), HW4 (M = 90.85%, SD = 17.01), and HW5 (M = 97.68%, SD = 6.34). At the semester level, Semester 251 achieved M = 94.61% (SD = 7.09, N = 82), representing a decline of 4.99 percentage points from Semester 242's near-ceiling performance (Hedges' g = -0.90, 95% CI [-1.23, -0.57]). However, this decline must be interpreted in light of the substantially increased difficulty: doubling the number of items while eliminating retakes and requiring proctored administration would be expected to reduce scores. The more meaningful comparison is against the baseline.

Overall Comparison: Semesters 241 to 251.

Comparing the baseline (Semester 241) to the final implementation (Semester 251), homework performance improved by 18.61 percentage points (from M = 76.00% to M = 94.61%), yielding a very large effect size (Hedges' g = 1.20, 95% CI [0.83, 1.57]). This substantial gain occurred despite Semester 251's more challenging assessment conditions, suggesting that the structured workflow successfully supported deep procedural learning. The progressive reduction in standard deviation across semesters (241: SD = 20.61 → 242: SD = 2.47 → 251: SD = 7.09) further indicates that the AI-supported system consistently reduced achievement gaps.

Interpretation.

These two distinct patterns, dramatic gains under low-stakes conditions (Semester 242) and sustained performance under substantially stricter conditions (Semester 251), suggest genuine learning rather than superficial quiz optimization. Table 2 presents the complete descriptive statistics and effect sizes across all three semesters.

Table 2: Homework and Exam Performance Across Three Semesters

Measure	**Sem. 241 M (SD)**	**Sem. 242 M (SD)**	**Sem. 251 M (SD)**		**Effect Size**
Homework (%)	76.00	99.60	94.61		
N = 67/82/82	(20.61)	(2.47)	(7.09)		
				241→242	g = 1.70
				241→251	g = 1.20
				242→251	g = -0.90

Measure	Sem. 241 M (SD)	Sem. 242 M (SD)	Sem. 251 M (SD)		Effect Size
Midterm (%)	—	64.29	85.01		
N = —/82/82		(20.75)	(13.73)		
				242→251	g = 1.19
Final (%)	61.36	64.27	70.46		
N = 66/82/82	(17.17)	(16.04)	(17.83)		
				241→242	g = 0.17
				241→251	g = 0.52
				242→251	g = 0.36

Note. All effect sizes (Hedges' g) were significant at p < .001 except for the 241→242 final exam (p = .10). The 242→251 (g = 0.36, p < .01) and 241→251 (g = 0.52, p < .001) final exam comparisons are both significant, demonstrating moderate gains in comprehensive assessment performance. Semester 251 used stricter quiz conditions (10-item, 1-attempt, proctored) vs. *241/242 (5-item, 3-attempt, unproctored). Analysis of multiple-choice items (assessing factual knowledge) showed comparable effect sizes to overall performance (242→251: g = 0.28 for MCQ* vs. *g = 0.36 overall), indicating that the workflow supported both knowledge acquisition and application.*

Midterm Exam Performance (Targeted Knowledge Gain)

The proctored midterm exam assessed the transfer of procedural and conceptual knowledge to a high-stakes, closed-book environment. The exam consisted of 10 multiple-choice questions covering research terminology and APA citation rules plus 5 application-based questions requiring students to identify and correct citation errors, for a total of 15 items worth 20 points.

Midterm data were available for Semesters 242 and 251 (Semester 241 did not administer the midterm via ExamSoft). Semester 251 (M = 85.01%, SD = 13.73, N = 82) substantially outperformed Semester 242 (M = 64.29%, SD = 20.75, N = 82), representing a gain of 20.72 percentage points. This difference constitutes a large effect size (Hedges' g = 1.19, 95% CI [0.82, 1.56]).

Interpretation.

The large midterm gain observed from Semester 242 to Semester 251 provides evidence that the structured homework workflow supported genuine, deep learning instead of superficial quiz performance. Despite Semester 242

students having access to the same AI tools (CustomGPT and recap videos) and achieving near-ceiling homework scores, their midterm performance was only moderate. In contrast, Semester 251 students—who were required to complete the Guide → CustomGPT → Quizlet → Examplify sequence before each homework quiz—demonstrated substantially higher midterm performance. This pattern suggests that mandatory structured practice, rather than tool availability alone, was the key mechanism for the deep encoding of APA procedures and research terminology.

Additionally, Semester 251 exhibited both higher mean performance and lower variability (SD = 13.73 vs. 20.75), indicating more consistent mastery across the cohort. This reduction in variability aligns with the homework pattern and reinforces the interpretation that the structured workflow served as an equity tool, helping weaker students develop procedural fluency alongside their stronger peers.

Final Exam Performance (Comprehensive Assessment)

The proctored final exam assessed comprehensive knowledge of all course content, including research terminology, APA citation rules, research methods, and ethical conduct in research. The exam structure paralleled the midterm but covered the full semester's material, serving as a summative assessment of students' research writing competency.

Final exam data were available for all three semesters. Semester 251 (M = 70.46%, SD = 17.83, N = 82) outperformed both Semester 242 (M = 64.27%, SD = 16.04, N = 82) and baseline Semester 241 (M = 61.36%, SD = 17.17, N = 66). The gain over Semester 242 was 6.19 percentage points (Hedges' g = 0.36, 95% CI [0.05, 0.67], $p < .01$), and the gain over baseline was 9.10 percentage points (Hedges' g = 0.52, 95% CI [0.19, 0.85], $p < .001$).

Interpretation.

The final exam results provide additional evidence of transfer beyond the immediate practice context. While the effect size was moderate (g = 0.36 for 242→251) compared to the large midterm effect (g = 1.19), this pattern is expected given the final exam's comprehensive scope covering all semester content rather than just the material directly practiced in the structured workflow. The +6.19 percentage point improvement over Semester 242 suggests that the structured workflow supported retention and application of knowledge beyond specific content practiced during weekly homework cycles. The +9.10 percentage point gain over baseline (g = 0.52) demonstrates substantial, sustained improvement in comprehensive research writing knowledge across the full tri-semester intervention period. The progressive improvement across all three semesters confirms cumulative learning gains that extended to comprehensive, proctored assessments, indicating educationally meaningful transfer across diverse content areas.

Knowledge-Based Performance (Multiple-Choice Component)

To assess whether the structured workflow specifically enhanced factual knowledge acquisition, we analyzed performance on the multiple-choice component of the final exams. Semesters 241 and 242 included 10 multiple-choice items assessing research terminology and APA citation rules, while Semester 251 included 5 such items. For comparability, performance was calculated as percentage correct.

Semester 251 ($M = 77.56\%$, $SD = 26.74$, $N = 82$) outperformed both Semester 242 ($M = 71.10\%$, $SD = 17.78$, $N = 82$; Hedges' $g = 0.28$, $p < .05$) and baseline Semester 241 ($M = 68.94\%$, $SD = 17.46$, $N = 66$; Hedges' $g = 0.36$, $p < .01$). The gain over Semester 242 was 6.46 percentage points.

Interpretation.

The effect size for knowledge-based items ($g = 0.28$) was comparable to the overall exam ($g = 0.36$). This suggests that the workflow supported both factual knowledge and applied reasoning. Similar gains indicate that the system promoted comprehensive learning rather than privileging one type of knowledge over another.

Pass rates (≥60%, or 24/40 points) improved progressively across semesters. Semester 241 achieved 78.8% (52/66 students), semester 242 reached 84.3% (69/82 students), and semester 251 attained 85.4% (70/82 students). This represents a 6.6 percentage point increase from baseline to full implementation, demonstrating that the AI-supported intervention reduced failure rates while raising mean performance—achieving both excellence and equity goals.

Platform engagement and microlearning utility

Platform usage confirmed high and sustained engagement with AI support across the spring and fall 2025 semesters. The CustomGPT logged over 400 unique conversations during the two terms of its availability (Spring 2025 through December 9, 2025), with the most frequent queries relating to citation format and outline structure verification. Usage grew from approximately 100 conversations in late October to 400+ by early December, indicating sustained adoption throughout the fall semester. This high usage supports the idea of perceived usefulness (Davis, 1989), suggesting that the tool addressed genuine, high-need student learning gaps outside of class time.

YouTube analytics for the brief, focused recap videos indicated strong alignment with cognitive load theory. The 12 active videos (two obsolete videos removed due to curriculum restructuring from annotated bibliography to the literature review) generated 659 total views with 13.86 hours of cumulative watch time during Semester 251. The average retention rate was 39.51% across the series, with all views classified as engaged views (100% engagement rate). Playlist viewing showed deadline-driven behavior, with 77% of playlist views (41 of 53) occurring on a single day (December 3, 2025), likely corresponding to a

major assignment deadline. Students reported that the segmented videos were an essential component of the workflow, using them for just-in-time review prior to the CustomGPT practice stage. Semester 251 students also used NotebookLM for AS1 source comprehension; informal feedback indicated that this helped students understand scholarly texts before drafting. However, systematic usage data for NotebookLM were not collected.

Table 3: Student Perceptions of AI Integration and Responsible Use

Survey Item	M	SD
CustomGPT Utility		
CustomGPT helped me understand APA citation	6.23	0.94
CustomGPT saved me time on homework	6.08	1.02
CustomGPT improved the quality of my work	5.89	1.15
Academic Integrity Awareness		
I use AI as a coach, not to write for me	6.15	1.08
I understand when AI use is/is not acceptable	5.92	1.21
Linguistic Support (EMI Context)		
Arabic explanations helped me understand concepts	5.08	1.45
Workflow Components		
Video recaps were essential to my learning	5.87	1.09
Quizlet practice prepared me for proctored quizzes	6.01	1.03

Note. Semester 251: N = 39. Responses on a 7-point Likert scale (1 = Strongly Disagree, 7 = Strongly Agree). Response rate = 47.6% (39/82). High means (M > 5.0) across all items indicate strong student endorsement of the AI-supported workflow and responsible use framework. EMI = English-Medium Instruction.

The Semester 251 survey data strongly endorsed CustomGPT's utility and confirmed responsible use within the integrity-by-design framework (Table 3). All survey items received mean agreement scores above 5.0 on a 7-point Likert scale (1 = Strongly Disagree, 7 = Strongly Agree), with particularly high ratings for CustomGPT's role in understanding APA citation (M = 6.23, SD = 0.94) and for students' self-reported use of AI as a coach rather than ghostwriter (M = 6.15, SD = 1.08).

These findings show that AI was viewed as a positive, rule-enforcing coach, suggesting that pedagogical framing successfully mitigated academic integrity concerns by focusing on process evidence (work logs, drafts) and high-value scaffolding tasks. The moderate but significant uptake of the Arabic explanation feature (M = 5.08, SD = 1.45) also confirms the utility of providing optional L1 support for procedural terminology in an EMI context.

DISCUSSION

Effectiveness of Scaffolding and Targeted AI Support

The results demonstrate that AI-supported scaffolding can substantially improve academic performance in demanding EMI research writing courses, with effects varying by implementation approach. Semester 242's introduction of CustomGPT tutoring and recap videos produced dramatic gains in homework performance (+23.6 percentage points over baseline, $g \approx 1.70$), demonstrating strong tool effectiveness under low-stakes, multiattempt conditions. Semester 251 maintained high performance (+18.6 points over baseline, $g \approx 1.20$) despite substantially increased assessment difficulty (10 items vs. 5, single attempt vs. three attempts, proctored vs. unproctored), suggesting that the structured workflow successfully managed cognitive load while deepening learning (Mayer, 2009). By assigning the CustomGPT the role of immediate feedback provider for mechanical tasks (e.g., APA formatting, structure), the instructor was able to dedicate class time to higher-order skills, such as argumentation, critical source integration, and rhetorical awareness. This division of labor aligns with the literature advocating for explicit, scaffolded supports for multilingual writers in the Saudi EMI context (Elyas & Al-Hoorie, 2024). Consistent with Bayati et al.'s (2025) findings that international students rely on technology as a critical coping mechanism for academic adjustment, this study suggests that GenAI can function as an equitable "writing coach" that supports student resilience rather than merely acting as a shortcut.

Final exam performance demonstrated continued improvement across the intervention period. Semester 251 (M = 70.46%, SD = 17.83, N = 82) outperformed Semester 242 (M = 64.27%, SD = 16.04, N = 82; Hedges' g = 0.36, $p < .01$) and baseline Semester 241 (M = 61.36%, SD = 17.17, N = 66; Hedges' g = 0.52, $p < .001$), representing gains of +6.19 and +9.10 percentage points, respectively. While these effect sizes are moderate compared to the large midterm improvement (g = 1.19), this pattern is theoretically coherent: the final exam assessed comprehensive knowledge across all course content, whereas the midterm targeted specific material that students practiced intensively through the structured workflow. The sustained gains in the comprehensive final suggest that the workflow supported not only the memorization of isolated facts but also the deep encoding and transferable understanding of research writing principles (Mayer, 2009; Wang & Fan, 2025).

Analysis of knowledge-based test items further highlights the workflow's mechanisms. Multiple-choice performance (assessing factual knowledge of

research terminology and APA rules) improved with an effect size (g = 0.28) comparable to overall exam performance (g = 0.36), indicating that the intervention supported both declarative knowledge and applied skills. The slightly smaller MCQ effect suggests that the workflow may have been particularly effective for complex, application-based essay questions requiring the integration of multiple concepts. This pattern aligns with the pedagogical design: the Guide → CustomGPT → Quizlet → Examplify sequence emphasized understanding and application over rote memorization (Sweller, 1988; Mayer, 2009) but still produced strong gains in factual knowledge, indicating broad rather than narrow learning effects (Wang & Fan, 2025; Luo et al., 2025).

However, two observations counter the view that retrieval practice alone drove these gains: Quizlet alone cannot explain the strong survey endorsement of CustomGPT utility (M = 6.23 for APA understanding), and the workflow was designed so that CustomGPT provided conceptual explanations, while Quizlet reinforced terminology through retrieval practice. The combination, understanding followed by practice, likely produced the observed effects, consistent with research on interleaved conceptual and procedural learning (Mayer, 2009; Muali & Karlina, 2025; Wang & Fan, 2025).

Tool availability and performance equity (Semester 242)

The Semester 242 findings challenge assumptions that tool availability alone is insufficient for learning gains. With only the CustomGPT and recap videos introduced and no changes to assessment structure, homework performance improved dramatically (M = 99.60% vs. 241's M = 76.00%), with drastically reduced variability (SD = 2.47 vs. 20.61). This near-ceiling performance and minimal variation indicate that students successfully used AI support to master APA mechanics and research concepts at the quiz level (Wang & Fan, 2025). However, the moderate midterm performance in Semester 242 (M = 64.29%) compared to homework suggests a limitation: students could access and apply correct information during quizzes with AI support, but this access did not automatically translate to independent recall and application under exam conditions (Luo et al., 2025). The tools provided scaffolding but did not ensure deep encoding (Mayer, 2009).

The progressive reduction in homework variability across the three semesters provides evidence of the AI system's equity function (Bayati et al., 2025). Standard deviations decreased substantially: Semester 241 (SD = 20.61), Semester 242 (SD = 2.47), Semester 251 (SD = 7.09). The initial dramatic reduction (241 to 242) indicates that AI tool availability immediately helped lower-performing students achieve mastery of quiz content. The modest increase in variability from 242 to 251 (2.47 to 7.09) reflects the introduction of stricter conditions: single-attempt, proctored quizzes, which naturally produce greater differentiation. However, Semester 251's SD remained far below the baseline (7.09 vs. 20.61), confirming that the structured workflow sustained equity benefits while supporting deeper learning (Elyas & Al-Hoorie, 2024).

The pattern across semesters clarifies how different implementations achieve equity. Semester 242's AI tools provided on-demand support that helped all students achieve near-ceiling quiz performance, effectively eliminating the achievement gap for homework; however, this did not ensure deep learning, as evidenced by moderate midterm scores. Semester 251's mandatory workflow, in contrast, maintained equity (SD = 7.09, still much lower than the baseline's 20.61) while supporting transfer to proctored exams (Wang & Fan, 2025; Luo et al., 2025). This suggests that combining tool availability with structured practice produces both equity and genuine learning outcomes in EMI contexts where achievement gaps persist (Elyas & Al-Hoorie, 2024; Macaro et al., 2018).

The qualitative data highlight CustomGPT's critical role as a linguistic scaffold within this English-medium instruction (EMI) environment. The optional feature allowing students to request Arabic explanations for complex concepts received a mean agreement score of 5.08/7, indicating that it was highly valued. This finding connects directly to known challenges in EMI contexts where multilingual learners often face simultaneous cognitive load from both language acquisition and content mastery (Elyas & Al-Hoorie, 2024; Macaro et al., 2018). Language barriers in particular have been identified as a significant factor shaping how international and nonnative English-speaking students engage with and adopt new academic technologies (Ittefaq et al., 2025; Bayati et al., 2025). By offering a low-cost, just-in-time linguistic safety net, the CustomGPT allowed students to instantly confirm their understanding of complex procedural or disciplinary terms in their L1 (Arabic) before applying them in English. This capability likely reduced the intrinsic cognitive load associated with the foreign-language medium (Duke, 2026; Sweller, 1988; Mayer, 2009), enabling students to dedicate more working memory to the actual research-writing task.

Reframing Academic Integrity: AI as Coach, Not Ghostwriter

The most significant qualitative finding is the preliminary endorsement by students of the CustomGPT as a "coach, not ghostwriter." This perception is critical to the success of an AI-integrated curriculum, and it links directly to the Technology Acceptance Model (TAM). The design strategy focused on promoting the AI's perceived usefulness by limiting its scope to low-stakes, high-need tasks (checking citations, generating search strings, refining outlines), consistent with evidence that clearly defined AI roles increase student confidence and ethical engagement (Davis, 1989; Qian et al., 2025).

For instance, students commonly expressed that the tool was useful for managing the "mechanics" of research writing. One student stated, "CustomGPT helped me check citation formats and create proper APA references. It made referencing much easier and less stressful. [AI tools] saved time on editing and research, allowing me to focus on creativity" (S-251-Email-06).

The perception of AI as a nonthreatening, always available guide fostered a positive relationship with the tool, leading to the "coach" designation. Another student noted, "Not plagiarizing, i.e., copying what the AI says rather use it as a framework for the rest of the work" (S-242-Email-01), demonstrating that

students understood the AI's role as a structural guide rather than a content generator.

This explicit boundary, enforced by process-evidence assignments, created a trust relationship where students viewed the tool as an extension of the course's scaffolding rather than a means of circumventing the learning process (Luo et al., 2025; Ittefaq et al., 2025). This finding contrasts with earlier academic concerns about wholesale misuse (Mollick & Mollick, 2024), demonstrating that intentional pedagogical design can successfully reframe student interaction with GenAI from a tool for cheating to a personalized, anytime tutor. By enforcing a process-evidence paradigm (Liang et al., 2023) and structuring the AI's role as a guide, the course design effectively mitigated the risk of nonethical use while maximizing its utility for targeted skill acquisition (Qian et al., 2025).

Implications for International and EMI Student Support

The tri-semester pattern of results suggests that structured AI workflows can function as an "integrity-by-design" infrastructure for international and EMI students. Rather than treating generative AI as an external risk to be policed, the Research Writing Techniques workflow embedded AI at clearly defined points in the writing process—generating search strings, checking APA citations, and rehearsing outlines—while preserving human ownership of ideas and final language. For EMI instructors and program coordinators, this reframing offers a practical model for using GenAI to reduce cognitive load, stabilize performance across cohorts, and support multilingual learners without compromising academic standards.

For student support services, the findings underscore the importance of coordinated implementation across writing courses, writing centers, and academic skills units. A course-specific CustomGPT, aligned with local policies and trained on vetted materials, can be paired with short recap videos and low-stakes quizzes to create a coherent support structure around high-stakes assignments. Writing centers and international student offices could adapt this model by codesigning prompt templates for responsible AI use, integrating process-evidence requirements into consultations, and offering "AI as coach, not ghostwriter" workshops for students and faculty.

Finally, the results carry equity implications for EMI programs that enroll both domestic and international students. In this study, the structured workflow narrowed performance gaps while improving pass rates and scores on knowledge-based exam items, suggesting that multilingual learners benefit when access to AI is organized, transparent, and assessment-aligned rather than ad hoc. EMI programs beyond Saudi Arabia can adapt the underlying principles—segmented supports, explicit integrity boundaries, and optional L1 scaffolding for complex procedural concepts—to their own linguistic and policy contexts, using GenAI not only to manage risk but also to promote fairness and academic persistence among multilingual learners.

CONCLUSION

This tri-semester case study examined the staged integration of generative AI into research writing techniques at a private English-medium university in Saudi Arabia. The findings demonstrate that a coherent package of CustomGPT tutoring, brief recap videos, and a structured homework workflow can substantially improve research writing outcomes, reduce achievement gaps, and foster responsible AI use among first-year Arabic L1 students studying through English-medium instruction.

The progression from instructor-only AI use (Semester 241) to optional student-facing tools (Semester 242) to a mandatory structured workflow (Semester 251) revealed two distinct mechanisms: tool availability alone produced dramatic performance gains under low-stakes conditions, while the structured workflow maintained high performance under substantially more demanding assessment conditions and supported genuine knowledge transfer to comprehensive proctored exams. The large midterm effect ($g = 1.19$) and sustained final exam gains ($g = 0.52$ over baseline) suggest that the workflow promoted deep encoding rather than superficial quiz optimization.

Critically, the quantitative gains are mirrored and explained by the qualitative data. Students did not merely perform better — they articulated a coherent understanding of AI's appropriate role, endorsing the CustomGPT as a coach rather than a ghostwriter, and voluntarily using the Arabic explanation feature to reduce the language burden of EMI instruction. This convergence of performance data and student voice suggests that the observed gains were not the product of optimized test-taking behavior but of genuine shifts in how students approached research writing tasks.

Together, these findings point toward a broader principle for EMI programs serving multilingual learners: structured AI integration, when designed around process evidence and clear integrity boundaries, can function simultaneously as a pedagogical scaffold, an equity mechanism, and an academic integrity framework. As GenAI tools become normalized in higher education, the challenge is no longer whether to integrate them but how to do so in ways that deepen rather than displace learning. This study offers one evidence-based answer to that question, grounded in three semesters of classroom data from an underrepresented context in the empirical literature.

Limitations and Future Work

Several limitations should be noted. First, the quasiexperimental design lacks random assignment, limiting causal claims. While the sequential improvements strongly suggest treatment effects, alternative explanations (e.g., instructor refinement over time, cohort differences) cannot be entirely ruled out. Second, the study examines a single course at one institution, limiting generalizability to other EMI contexts, disciplinary areas, and institutional settings. Replication across diverse contexts is needed to establish the robustness of these findings. Third, the 47.6% survey response rate, while adequate, may introduce self-selection bias in

the perception data. Fourth, Semester 251 introduced a curricular change (Assignment 1 shifted from Annotated Bibliography to Literature Review) alongside the structured workflow, making it difficult to isolate the contribution of each modification. However, final exam content remained identical across all semesters, and the Literature Review's emphasis on synthesis complements rather than confounds the AI workflow's focus on mechanical accuracy.

Future research should include several directions to build on these preliminary findings. First, a longitudinal extension of this study across subsequent semesters (252, 261, and beyond) is needed to reveal whether the observed effects persist, diminish, or strengthen as AI tools become normalized in the curriculum. Second, replication across multiple courses and institutions in the Gulf region would establish the generalizability of these findings to diverse English-medium instruction (EMI) populations and disciplinary contexts. Third, linking individual CustomGPT session frequency and depth to performance outcomes could reveal optimal usage thresholds and identify students needing additional scaffolding. Additionally, future research should isolate the effects of different AI tool types—distinguishing "reading assistants" (e.g., NotebookLM for source comprehension) from "writing tutors" (e.g., CustomGPT for citation and structure)—to identify which scaffolds contribute most to learning gains. Finally, controlled experimental designs comparing structured AI workflows to traditional instruction would strengthen causal inferences and isolate the specific mechanisms driving the observed gains in procedural knowledge acquisition and equity outcomes.

TRANSPARENCY AND DISCLOSURES

AI Use and Acknowledgment

Generative AI tools were used in limited ways to support manuscript preparation: (a) drafting alternative phrasings for section headings and transitions; (b) checking grammar and clarity at the sentence level; and (c) generating boilerplate options for survey/consent wording that were subsequently customized. The author independently designed the study, adapted or created all teaching materials and instruments, collected and cleaned the data, conducted all analyses, and interpreted the findings. The author verified every reference and fact, edited all AI-suggested text, and assumed full responsibility for the final content. No student-identifiable data or unpublished exam materials were uploaded to public AI services, and no AI systems were used to generate, alter, or fabricate data. This disclosure aligns with COPE guidelines and the Journal of International Students requirements.

Acknowledgments

The author gratefully acknowledges the institutional support of Prince Sultan University, the guidance of the Research Writing Techniques Course Group Supervisor and PSU colleagues, and the students who volunteered feedback and interviews.

REFERENCES

Bayati, N., Denson, C., & Asunda, P. (2025). Examining barriers and facilitators of graduate international students in the U.S. *Journal of International Students, 15*(5), 139–158. https://doi.org/10.32674/b6bgby16

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

Duke, B. (2026). Artificial intelligence for inclusive and sustainable international education: A critical conceptual review. *Journal of International Students, 16*(10), 167–198. https://doi.org/10.32674/96t36095

Elyas, T., & Al-Hoorie, A. H. (2024). English-medium instruction in higher education in Saudi Arabia. In K. Murata (Ed.), *The Routledge handbook of English-medium instruction in higher education* (pp. 285–298). Routledge. https://doi.org/10.4324/9781003011644-22

Freeman, J., & Higher Education Policy Institute. (2025, February). The nature of student use of generative AI: Findings from the 2025 Student Academic Experience Survey (Policy Note 61). HEPI. https://www.hepi.ac.uk/wp-content/uploads/2025/02/HEPI-Kortext-Student-Generative-AI-Survey-2025.pdf

Guo, P. J., Kim, J., & Rubin, R. (2014). How video production affects student engagement: An empirical study of MOOC videos. In *Proceedings of the First ACM Conference on Learning at Scale* (pp. 41–50). https://doi.org/10.1145/2556325.2566239

Institute of International Education. (2024). Open Doors report on international educational exchange. IIE. https://opendoorsdata.org/

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

Liang, W., Chu, X., Li, Z., Huang, Z., Xue, K., Nguyen, T. D., Zhang, C., Zhu, S.-C., & Li, Y. (2023). GPT detectors are biased against non-native English writers. *Patterns*, 4(8), 100779. https://doi.org/10.1016/j.patter.2023.100779

Luo, J., Zheng, C., Yin, J., & Teo, H. H. (2025). Design and assessment of AI-based learning tools in higher education: a systematic review. *International Journal of Educational Technology in Higher Education, 22*(1), 42. https://doi.org/10.1186/s41239-025-00540-2

Macaro, E., Curle, S., Pun, J., An, J., & Dearden, J. (2018). A systematic review of English-medium instruction in higher education. *Language Teaching*, 51(1), 36–76. https://doi.org/10.1017/S0261444817000350

Mayer, R. E. (2009). *Multimedia learning* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511811678

Mogavi, R., You, L., Zhao, J., & Cai, R. (2024). Exploring the influence of ChatGPT on student academic performance: A systematic review and meta-analysis. *Education and Information Technologies*. Advance online publication. https://doi.org/10.1007/s10639-024-13148-2

Mollick, E. R., & Mollick, L. (2024). *Instructors as innovators: A future-focused approach to new AI learning opportunities, with prompts* (SSRN Working Paper No. 4802463). https://doi.org/10.2139/ssrn.4802463

Muali, C., & Karlina, L. (2025). The effect of microlearning integration in digital platforms on student engagement: An experimental study in higher education. *Journal of Education Technology, 9*(1), 21–30. https://doi.org/10.23887/jet.v9i1.92613

Qian, Y., Wang, X., & Zhang, C. (2025). Pedagogical applications of generative AI in higher education: A systematic literature review. *TechTrends*. Advance online publication. https://doi.org/10.1007/s11528-025-01100-1

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4

UNESCO. (2023). *Global education monitoring report 2023: Technology in education: A tool on whose terms?* https://www.unesco.org/gem-report/en/technology

Wang, J., & Fan, W. (2025). The effect of ChatGPT on students 'learning performance, learning perception, and higher-order thinking: Insights from a meta-analysis. *Humanities and Social Sciences Communications, 12*, 621. https://doi.org/10.1057/s41599-025-04787-y

Author bio

Erick C.W. Nelson is a Lecturer in the Linguistics and Translation Department at Prince Sultan University, Saudi Arabia. He has over 25 years of experience in English language teaching and curriculum design. His major research interests lie in the area of integrating generative AI into research writing, English-medium instruction, and academic integrity for Arabic L1 students. Email: enelson@psu.edu.sa