OPEN ACCESS
Nusantara: Indonesian Journal of Islamic Studies
ISSN: 2809-283X (e); 2809-2902 (p)

Vol. 6, No. 2 (2026), pp. 34-45
doi: 10.54471/nusantara
https://ejournal.nun.or.id/index.php/nusantara

Generative AI as a Writing Feedback Tool in Arabic Language Learning: A Mixed-Methods Study in Indonesian Islamic Higher Education

Himmatul Maulidiyah[1], Khusnul Khotimah[2], Khoirun Nisa'[3], Mufidah Sirodjul Ummah[4], Isro'ul Laili[5]*

[1]Department of Arabic Language Education, Institut Agama Islam Al Yasini Pasuruan, Indonesia
*Corresponding author: isro.pba.alyasini@gmail.com

Received: 2026-05-29; Approved: 2026-06-27; Published: 2026-07-16

ABSTRACT

Background and research gap Generative artificial intelligence (GenAI) is rapidly entering second-language writing classrooms, yet empirical evidence comes almost entirely from English. Its reliability and pedagogical value for Arabic, a morphologically rich language taught in Indonesia as both a religious and an academic language, remain underexplored. Purpose This study examines whether GenAI-generated feedback improves students' Arabic writing performance, how linguistically accurate that feedback is, and how students perceive and engage with it. Design/methodology/approach A convergent parallel mixed-methods design combined a quasi-experiment with 62 Arabic Language Education students (experimental n = 31; control n = 31) over eight weeks, expert accuracy coding of 1,248 feedback points, a perception questionnaire, and semi-structured interviews with 12 students. Key Findings The experimental group outperformed the control group on post-test writing scores ($F(1, 59) = 14.62$, $p < .001$, partial $\eta^2 = .20$). GenAI feedback was 81.3% accurate overall but weakest on iʿrāb (case endings). Students valued its immediacy and low-anxiety tone while adopting "trust but verify" strategies. Research limitations/implications The single-institution sample and short intervention limit generalisability; GenAI feedback should complement, not replace, lecturer feedback. Originality/value The study provides early mixed-methods evidence on GenAI feedback for Arabic writing and proposes a lecturer-mediated feedback model for Islamic higher education.

KEYWORDS: generative artificial intelligence, Arabic writing, automated written feedback, mixed methods, Islamic higher education

© 20xx The Author(s). Published by CV. Murta Media Karya Surabaya, Management by Nun Academic Solutions, Lumajang
This is an Open Access article distributed under the terms of the is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (http://creativecommons.org/licenses/by-sa/4.0/)

INTRODUCTION

Writing (mahārah al-kitābah) is widely regarded as the most demanding of the four Arabic language skills for Indonesian learners because it requires simultaneous control of morphology (ṣarf), syntax (naḥw), orthography (imlāʾ), and rhetorical convention (uslūb). In Arabic Language Education (Pendidikan Bahasa Arab, PBA) programmes at Islamic higher education institutions, students are expected to produce academic and functional texts in Modern Standard Arabic, yet large classes and heavy teaching loads mean that lecturers can rarely provide timely, individualised written corrective feedback. Feedback that arrives late, or that is limited to surface corrections, does little to promote the behavioural, cognitive, and affective engagement that drives revision and learning (Zhang & Hyland, 2022).

The public release of ChatGPT in late 2022 transformed this landscape. Large language models (LLMs) can now generate fluent explanations, detect errors, and suggest revisions on demand, prompting both enthusiasm and concern across higher education (Kasneci et al., 2023; Lo, 2023; Farrokhnia et al., 2024). In second-language (L2) writing, studies report that generative artificial intelligence (GenAI) can support idea generation, language accuracy, and revision (Barrot, 2023; Su et al., 2023; Yan, 2023), and that AI-generated feedback can be as effective as human feedback for some outcomes (Escalante et al., 2023; Steiss et al., 2024). Other studies caution that such feedback may be generic, occasionally inaccurate, and liable to encourage over-reliance or academic dishonesty (Cotton et al., 2024; Tlili et al., 2023).

This research base, however, is overwhelmingly anchored in English as a foreign language (EFL). Arabic differs from English in ways that matter for automated feedback: it is built on a root-and-pattern morphology, marks grammatical case and mood through short vowels that are usually omitted in writing, and exhibits diglossia between Modern Standard Arabic and colloquial varieties. Emerging work on Arabic instruction in Indonesia suggests that lecturers see promise in ChatGPT for generating materials and models (Nasaruddin, 2024), while studies among Arab learners writing in English highlight both appreciation of and scepticism towards AI-based evaluation (Al-Zubaidi et al., 2024; Alshayban, 2026). What remains missing is rigorous evidence on how accurate GenAI feedback is when the target language is Arabic itself, whether it improves Arabic writing outcomes, and how learners in religious educational settings interpret it. Indonesian Islamic higher education offers a distinctive setting for this inquiry. Many PBA students enter university from pesantren backgrounds in which Arabic grammar is learned through classical texts and in which the authority of the teacher (ustādh) carries moral as well as academic weight. How students reconcile machine-generated corrections with this tradition, and with the Islamic value of honesty (amānah) in scholarship, has not yet been examined.

Accordingly, this study addresses three research questions: (RQ1) Does GenAI-assisted feedback improve the Arabic writing performance of PBA students compared with lecturer-only feedback? (RQ2) How accurate is GenAI feedback across different categories of Arabic linguistic features? (RQ3) How do students perceive, use, and negotiate GenAI feedback in their writing and revision practices? By integrating quantitative and qualitative evidence, the study offers one of the first mixed-methods accounts of GenAI writing feedback in Arabic within PTKI and proposes a practical, lecturer-mediated model for its responsible use.

Literature Review

Feedback is among the most powerful influences on learning, but its effect depends on whether learners engage with it (Zhang & Hyland, 2022). Earlier automated writing evaluation tools offered mostly form-focused corrections, whereas LLM-based tools can provide dialogic, content-

level, and explanatory feedback in natural language. Guo and Wang (2024) found that ChatGPT produced a larger volume of feedback than EFL teachers, covering content, organisation, and language, although teachers judged part of it less relevant. Steiss et al. (2024) reported that well-trained human evaluators still provided higher-quality feedback on most dimensions, yet ChatGPT's feedback was close enough to be valuable where human feedback is scarce. Mahapatra (2024), using a mixed-methods intervention, showed that ChatGPT feedback significantly improved ESL students' academic writing, and Teng (2024) described EFL learners treating ChatGPT as a companion in feedback-seeking rather than as a threat.

Accuracy is a central concern. Mizumoto and Eguchi (2023) demonstrated that GPT-based scoring of L2 English essays reached acceptable agreement with benchmark scores, particularly when combined with linguistic features, but cautioned against unsupervised use. Barrot (2023) and Kohnke et al. (2023) noted that LLMs may invent rules, over-correct stylistic choices, and homogenise learners' voices, so teachers need the digital competence to mediate their output. For Arabic, these risks are amplified because LLMs are trained on far less Arabic than English data and because grammatical accuracy depends heavily on case and mood endings that are seldom written. Studies in Arabic-speaking contexts reveal mixed attitudes: learners appreciate ChatGPT's speed and motivational value (Ali et al., 2023), whereas academics worry about authenticity and the erosion of critical thinking (Al-Zubaidi et al., 2024).

Students' willingness to adopt GenAI depends on perceived usefulness, ease of use, and trust (Strzelecki, 2024; Zou & Huang, 2023). Surveys in higher education show that students value personalised and immediate support but are concerned about accuracy, privacy, and over-reliance (Chan & Hu, 2023). Integrity concerns are especially salient: Cotton et al. (2024) argued that institutions should redesign assessment rather than rely on detection, and Werdiningsih et al. (2024) found Indonesian EFL students negotiating a balance between AI assistance and authorial authenticity. Jeon and Lee (2023) framed teachers and ChatGPT as complementary, with teachers orchestrating and evaluating AI contributions. Within Indonesian Arabic education, Nasaruddin (2024) mapped how lecturers can prompt ChatGPT for instructional needs but did not test its feedback empirically. The present study builds on these strands by examining Arabic writing feedback through combined performance, accuracy, and perception data, interpreted through an Islamic ethic of scholarly honesty.

METHODS

Research Design

This study adopted a convergent parallel mixed-methods design in which quantitative and qualitative data were collected concurrently, analysed separately, and merged during interpretation, an approach increasingly used to capture both the outcomes and the lived experience of AI-supported writing (Mahapatra, 2024). The quantitative strand used a quasi-experimental, non-equivalent pre-test post-test control group design to address RQ1, complemented by a content analysis of feedback accuracy (RQ2) and a perception questionnaire. The qualitative strand used semi-structured interviews to explain how students perceived and negotiated GenAI feedback (RQ3).

Setting and Participants

The study was conducted in the Arabic Language Education (PBA) Study Programme of Institut Agama Islam Al Yasini Pasuruan, East Java, a private PTKI affiliated with a pesantren. Participants were 62 fourth-semester students enrolled in two intact classes of the Insyāʾ (Arabic

composition) course during the odd semester of the 2025/2026 academic year. One class was randomly assigned as the experimental group (n = 31) and the other as the control group (n = 31). Participants were aged 19 22 years, and 71% were pesantren graduates. For the qualitative strand, 12 students from the experimental group were purposively selected to represent high, middle, and low pre-test achievement.

Intervention Procedures

Over eight weeks, both groups completed four writing tasks of 150 250 words (descriptive, narrative, expository, and argumentative) using a draft feedback revision cycle taught by the same lecturer. In the experimental group, students submitted each first draft to ChatGPT (GPT-4o) using a structured prompt developed by the research team. The prompt asked the model to identify errors in naḥw, ṣarf, imlāʾ, vocabulary, and organisation; to explain each problem in Indonesian with an example; and to refrain from rewriting the whole text. Students then revised their drafts and the lecturer provided brief confirmatory comments on the final version. Before the intervention, the experimental group attended a 90-minute feedback-literacy session on verifying AI suggestions against grammar references. In the control group, the lecturer provided conventional written corrective feedback on each first draft before revision.

Instruments

Four instruments were used. First, Arabic writing performance was measured with pre-test and post-test essays on parallel prompts, scored with an analytic rubric comprising content (25 points), organisation (20), vocabulary (20), grammar (25), and mechanics (10). Two trained raters scored all essays blind to group and test occasion, and inter-rater reliability was high (ICC = .89). Second, all GenAI feedback points generated on experimental-group drafts (N = 1,248) were coded by two Arabic linguistics experts as accurate, partially accurate, or inaccurate within five linguistic categories (Cohen's κ = .84); each point was also coded for student uptake in the revised draft. Third, a 24-item, five-point Likert questionnaire measured perceived usefulness, perceived ease of use, trust in accuracy, and concern about dependence and integrity, adapted from technology-acceptance studies of ChatGPT (Strzelecki, 2024; Zou & Huang, 2023); internal consistency was acceptable (Cronbach's α = .81 .90). Fourth, semi-structured interviews of 30 45 minutes were conducted in Indonesian, audio-recorded, and transcribed verbatim.

Data Analysis

Quantitative data were analysed using IBM SPSS Statistics 26. After confirming normality (Shapiro Wilk), homogeneity of variance (Levene's test), and homogeneity of regression slopes, a one-way analysis of covariance (ANCOVA) compared post-test scores between groups with pre-test scores as the covariate; partial eta squared (η^2) was reported as the effect size. Feedback accuracy and uptake were summarised as frequencies and percentages, and a chi-square test examined the association between accuracy and uptake. Interview data were analysed with reflexive thematic analysis (Braun & Clarke, 2021) assisted by NVivo 14, moving from familiarisation and coding to theme development and review. Trustworthiness was supported through member checking, peer debriefing among the five researchers, and methodological triangulation. The two

strands were integrated through a joint display that compared statistical results with qualitative themes.

Ethical Considerations

Participation was voluntary, and all students provided written informed consent. Students were assured that participation would not affect their course grades, and all data were anonymised. The study received approval from the Research and Community Service Institute (LP2M) of Institut Agama Islam Al Yasini Pasuruan.

RESULTS

The results are presented according to the three research questions: the effect of GenAI feedback on writing performance, the accuracy of the feedback, and students' perceptions and engagement.

Effect of GenAI Feedback on Arabic Writing Performance (RQ1)

Table 1 presents the descriptive statistics for both groups. The two groups had comparable pre-test scores, confirming their initial equivalence ($t(60) = 0.25$, $p = .806$). Both groups improved, but the experimental group's gain was considerably larger.

Table 1. Arabic writing scores by group

Group	Pre-test M (SD)	Post-test M (SD)	Adjusted post-test M (SE)
Experimental (GenAI + lecturer), n = 31	61.4 (8.2)	76.8 (7.1)	76.6 (1.2)
Control (lecturer only), n = 31	60.9 (7.6)	69.3 (7.9)	69.5 (1.2)
Difference	0.5	7.5	7.1

Note. Scores range from 0 to 100. Adjusted means are controlled for pre-test scores.

Assumption checks supported the use of ANCOVA: post-test scores were normally distributed in both groups (Shapiro Wilk $p > .05$), variances were homogeneous (Levene's $F(1, 60) = 0.42$, $p = .519$), and there was no significant group × pre-test interaction ($F(1, 58) = 0.87$, $p = .355$). After controlling for pre-test scores, the experimental group significantly outperformed the control group on the post-test ($F(1, 59) = 14.62$, $p < .001$, partial $\eta^2 = .20$), a large effect. Analysis of the rubric components showed that the advantage was greatest for mechanics (adjusted difference = 1.9 of 10 points) and vocabulary (2.1 of 20), moderate for grammar (1.8 of 25), and smallest for content and organisation (1.3 of 45 combined). GenAI feedback therefore produced its clearest gains in surface and lexical accuracy, while higher-order improvements were more modest.

Accuracy and Uptake of GenAI Feedback (RQ2)

ChatGPT generated 1,248 feedback points on 124 first drafts, an average of 10.1 points per draft. As shown in Table 2, 81.3% of the points were judged accurate, 10.5% partially accurate, and 8.2% inaccurate. Accuracy varied markedly by linguistic category. Feedback on orthography, such as hamzah placement, tāʾ marbūṭah, and alif maqṣūrah, was the most reliable, followed by vocabulary and morphology. Feedback on iʿrāb was the least reliable: about one in five comments on case and mood endings was wrong, typically because the model misidentified the governing element (ʿāmil) in nominal sentences, misapplied the rules of kāna and inna and their sisters, or treated a correct accusative ḥāl as an error.

Table 2. Accuracy of GenAI feedback by linguistic category (N = 1,248)

Category	Feedback points (n)	Accurate (%)	Partially accurate (%)	Inaccurate (%)
Orthography	286	92.3	4.9	2.8
Vocabulary and diction	271	85.6	9.2	5.2
Morphology	238	82.4	10.5	7.1
Syntax: case and mood endings	247	64.0	15.8	20.2
Organisation and coherence	206	80.1	13.6	6.3
Total	1,248	81.3	10.5	8.2

Note. Coded independently by two Arabic linguistics experts (Cohen's $\varkappa$ = .84); disagreements were resolved through discussion.

Student uptake was high but selective. Overall, 73.6% of feedback points were incorporated into revised drafts. Uptake was significantly associated with accuracy ($\chi^2(2, N = 1,248) = 97.36$, $p < .001$): students adopted 78.4% of accurate points but only 31.4% of inaccurate ones. Even so, 32 inaccurate suggestions, most concerning iʿrāb, were carried into final drafts, and the lecturer's confirmatory comments corrected 26 of them. This pattern indicates that students filtered many erroneous suggestions but that lecturer mediation remained necessary for the hardest grammatical cases.

Students' Perceptions and Engagement (RQ3)

Questionnaire results (Table 3) show that experimental-group students rated GenAI feedback as highly useful and easy to use. Trust in its accuracy was only moderate, and concern about dependence and academic integrity was relatively high, suggesting a positive but cautious stance.

Table 3. Students' perceptions of GenAI feedback (n = 31)

Dimension	Items	M	SD	α
Perceived usefulness	6	4.21	0.52	.90
Perceived ease of use	6	4.35	0.48	.86
Trust in feedback accuracy	6	3.42	0.71	.83
Concern about dependence and integrity	6	3.68	0.66	.81

Note. Five-point scale (1 = strongly disagree, 5 = strongly agree).

Reflexive thematic analysis of the interviews generated four themes that explain these patterns.

Theme 1: Immediate and face-saving feedback. Students consistently described GenAI feedback as fast, patient, and free from the embarrassment they associated with having errors exposed in front of a respected lecturer. One student explained, "I can ask the same question about mafʿūl bih five times and it never gets tired. With the ustādh I feel shy to ask again" (S4). Several students reported writing more drafts than usual because feedback was available at any time, including late at night in the pesantren dormitory.

Theme 2: Trust, but verify. Students did not accept feedback uncritically. They described checking suggestions, especially on iʿrāb, against grammar references such as al-Naḥw al-Wāḍiḥ and Matn al-Ājurrūmiyyah, consulting peers, or regenerating the response. "When it says my ending

is wrong, I open my kitab first. Sometimes the AI is the one who is wrong" (S9). Higher-achieving students were more confident in rejecting incorrect suggestions, whereas lower-achieving students more often accepted them, a pattern consistent with the uptake of the 32 inaccurate points.

Theme 3: From correction to explanation. Students valued explanations more than corrections. Because the prompt asked for reasons and examples in Indonesian, several students said they finally understood rules they had memorised in the pesantren without applying them in writing. However, some admitted that when deadlines were close they copied corrections without reading the explanations, which limited learning.

Theme 4: Amānah and ownership of writing. Students framed responsible AI use in moral terms. They distinguished between using AI to "check" their work, which they considered permissible, and letting AI "write" for them, which they saw as a breach of amānah (trust and honesty). "If the sentences are from the AI, then it is not my work and not my knowledge. That is not barakah" (S2). Students welcomed clear guidelines from the programme on acceptable use.

Integration of Quantitative and Qualitative Findings

Merging the two strands produced three meta-inferences. First, the statistical gain in writing scores (RQ1) was explained by the immediacy of feedback and the increase in revision cycles described in Theme 1. Second, the moderate trust ratings and the accuracy gap for iʿrāb (RQ2) converged with the verification practices reported in Theme 2, confirming that students' scepticism was well founded. Third, the relatively small gains in content and organisation aligned with Theme 3: students used GenAI mainly for language-level correction, and its potential for higher-order feedback was underused. Taken together, the findings were complementary rather than contradictory.

DISCUSSION

This study investigated GenAI as a writing feedback tool in Arabic language learning at an Indonesian Islamic higher education institution. The findings show that GenAI feedback, when combined with brief lecturer confirmation, improved students' Arabic writing performance more than lecturer-only feedback; that its accuracy was high overall but uneven across linguistic categories; and that students engaged with it enthusiastically yet critically, drawing on religious and pesantren values to regulate its use.

GenAI Feedback and Arabic Writing Performance

The large effect on post-test scores extends to Arabic the gains reported in EFL contexts (Mahapatra, 2024; Su et al., 2023; Yan, 2023). It also supports evidence that AI-generated feedback can approach the effectiveness of human feedback when human feedback is limited in frequency (Escalante et al., 2023; Steiss et al., 2024). The mechanism suggested by the qualitative data is not that GenAI feedback was superior in quality to the lecturer's, but that it was available immediately and repeatedly, allowing more revision cycles. This interpretation is consistent with the view that feedback works through learner engagement rather than through information alone (Zhang & Hyland, 2022), and with reports that ChatGPT can stimulate engagement in writing classes (Anjarani et al., 2024).

The concentration of gains in mechanics and vocabulary, with smaller improvements in content and organisation, mirrors earlier observations that students tend to use ChatGPT for language polishing rather than for developing ideas (Barrot, 2023; Guo & Wang, 2024). For Arabic writing pedagogy, this suggests that prompts and tasks should be redesigned to elicit feedback on argumentation, cohesion, and rhetorical conventions, not just on correctness. Lecturers might, for example, ask students to request feedback on the logical connection between paragraphs or on the use of cohesive devices such as fa-, thumma, and lākinna before requesting grammar correction.

The Accuracy Gap in Arabic Grammar

The most distinctive contribution of this study concerns feedback accuracy. An overall accuracy rate of about four in five is encouraging and broadly in line with studies showing reasonable agreement between LLM and human judgements of L2 writing (Mizumoto & Eguchi, 2023). Yet the sharp decline for iʿrāb reveals a language-specific weakness that EFL research cannot detect. Case and mood endings in Arabic are determined by syntactic relationships that may span several words and are usually not written, so the model must infer an abstract grammatical analysis rather than match surface patterns. The tendency of GenAI to produce confident but incorrect explanations, noted in general reviews (Kasneci et al., 2023; Lo, 2023; Dwivedi et al., 2023), is therefore especially consequential in the area of grammar that Indonesian Arabic curricula treat as foundational.

This finding has practical consequences. Because lower-achieving students were more likely to accept incorrect iʿrāb suggestions, unmediated GenAI feedback could reinforce errors among the learners who most need accurate guidance. The lecturer's confirmatory review, which corrected most of the inaccurate suggestions that reached final drafts, appears to be an essential safeguard. This supports the argument that teachers and GenAI are complementary, with teachers orchestrating and validating AI contributions (Jeon & Lee, 2023; Kohnke et al., 2023), and it aligns with calls for clear institutional guidance on GenAI use (Rudolph et al., 2023).

Critical Engagement and the Ethic of Amānah

Students' positive perceptions of usefulness and ease of use echo technology-acceptance research on ChatGPT in higher education (Strzelecki, 2024; Zou & Huang, 2023; Ngo, 2023), and the appreciation of low-anxiety, always-available feedback reflects findings among university students more broadly (Chan & Hu, 2023) and among Arab learners specifically (Ali et al., 2023; Alshayban, 2026). Students in this study were nonetheless less trusting than many EFL samples, which seems appropriate given the accuracy gap they experienced. Their habit of verifying suggestions against classical grammar texts shows that pesantren literacy practices can serve as a resource for critical AI literacy rather than an obstacle to it.

The theme of amānah adds a religious-ethical dimension to debates about AI and academic integrity. Whereas much of the literature frames integrity in terms of detection and assessment design (Cotton et al., 2024; Bašić et al., 2023), the students in this study framed it as a matter of personal honesty and of the blessing (barakah) of knowledge earned through effort. This internal moral framework resonates with the authenticity concerns of Indonesian EFL students (Werdiningsih et al., 2024) and with Arab academics' worries about authorship (Al-Zubaidi et al.,

2024), but it grounds them in an Islamic epistemology of learning. PTKI institutions can build on this framework by articulating AI-use policies in terms of amānah, distinguishing legitimate use for checking and learning from illegitimate use for content generation.

A Lecturer-Mediated Model for GenAI Feedback

Synthesising these findings, we propose a four-stage lecturer-mediated model for using GenAI feedback in Arabic writing courses. In the first stage, preparation, lecturers provide structured prompts and train students in feedback literacy, including how to verify suggestions against grammar references. In the second stage, AI-assisted drafting, students obtain explanatory feedback on their first drafts and revise them while keeping a record of accepted and rejected suggestions. In the third stage, lecturer validation, lecturers focus their limited time on higher-order issues and on the error categories in which GenAI is least reliable, particularly iʿrāb. In the fourth stage, reflection, students reflect on what they learned and on their use of AI in light of amānah. The model reallocates, rather than replaces, the lecturer's role, consistent with recommendations that AI in higher education should augment human teaching (Ouyang et al., 2022; Tlili et al., 2023; Farrokhnia et al., 2024).

Limitations and Future Research

Several limitations should be acknowledged. The study involved a single private PTKI and intact classes, so selection effects cannot be excluded and the findings may not generalise to state Islamic universities or to learners without pesantren backgrounds. The eight-week intervention did not allow assessment of long-term retention or transfer to new writing tasks. Only one GenAI model and one prompt design were examined, and model performance on Arabic is changing rapidly. Future research should use larger multi-site samples, delayed post-tests, and comparisons across models, including Arabic-centric LLMs, and should examine how different prompt designs affect the accuracy of feedback on iʿrāb and on higher-order features of Arabic writing (Abduljawad, 2024; Nasaruddin, 2024).

CONCLUSION

This mixed-methods study shows that GenAI can be a valuable writing feedback tool in Arabic language learning when it is integrated thoughtfully into instruction. Students who received GenAI feedback with lecturer confirmation achieved significantly higher Arabic writing scores than those who received lecturer feedback alone, with the strongest gains in mechanics and vocabulary. The feedback was accurate in most cases but notably unreliable for iʿrāb, the area of Arabic grammar that is most dependent on abstract syntactic analysis. Students welcomed the immediacy and low-anxiety character of GenAI feedback, verified its suggestions against classical grammar references, and regulated their use through the Islamic value of amānah.

For practice, PBA programmes in PTKI should adopt GenAI feedback as a complement to, not a substitute for, lecturer feedback; provide structured prompts and feedback-literacy training; direct lecturer attention to the error categories in which GenAI is weakest; and formulate AI-use guidelines grounded in amānah. For theory, the study demonstrates that the effectiveness of GenAI feedback is language-specific and that learners' religious and educational traditions shape how they engage with it. Further research across institutions, models, and longer time frames is

needed to refine the proposed lecturer-mediated model and to support the responsible use of GenAI in Arabic language education.

Author Contributions

Himmatul Maulidiyah: Conceptualisation, research design, and drafting of the original manuscript.
Khusnul Khotimah: Data collection, implementation of the intervention, and quantitative data analysis.
Khoirun Nisa': Coding of GenAI feedback accuracy and qualitative data analysis.
Mufidah Sirodjul Ummah: Literature review, interviews, and interpretation of findings.
Isro'ul Laili: Supervision, methodology validation, review and editing, and language proofreading.

Acknowledgements

The authors thank the Arabic Language Education Study Programme and the Research and Community Service Institute (LP2M) of Institut Agama Islam Al Yasini Pasuruan for supporting this study, the expert raters for coding the feedback data, and all students who participated.

Declaration of Competing Interest

The authors declare that there is no conflict of interest.

Ethical Approval

This study was approved by the Research and Community Service Institute (LP2M) of Institut Agama Islam Al Yasini Pasuruan. All participants gave written informed consent, and all data were anonymised.

REFERENCES

Abduljawad, S. A. (2024). Investigating the impact of ChatGPT as an AI tool on ESL writing: Prospects and challenges in Saudi Arabian higher education. *International Journal of Computer-Assisted Language Learning and Teaching, 14*(1), 1 19. https://doi.org/10.4018/IJCALLT.367276

Ali, J. K. M., Shamsan, M. A. A., Hezam, T. A., & Mohammed, A. A. Q. (2023). Impact of ChatGPT on learning motivation: Teachers and students' voices. *Journal of English Studies in Arabia Felix, 2*(1), 41 49. https://doi.org/10.56540/jesaf.v2i1.51

Alshayban, A. (2026). ChatGPT as a feedback provider: Arab students' perceptions of AI-based writing evaluation in academic English. *Ijaz Arabi: Journal of Arabic Learning, 9*(3), 1082 1098. https://doi.org/10.18860/ijazarabi.v9i3.42134

Al-Zubaidi, K., Jaafari, M., & Touzani, F. Z. (2024). Impact of ChatGPT on academic writing at Moroccan universities. *Arab World English Journal*(Special Issue on ChatGPT), 4 25. https://doi.org/10.24093/awej/ChatGPT.1

Anjarani, S., Furqon, M., & Rakhmawati, F. (2024). ChatGPT to foster students' engagement in writing class: An intervention study. *Voices of English Language Education Society, 8*(3), 630 640. https://doi.org/10.29408/veles.v8i3.27687

Barrot, J. S. (2023). Using ChatGPT for second language writing: Pitfalls and potentials. *Assessing Writing, 57*, 100745. https://doi.org/10.1016/j.asw.2023.100745

Bašić, Ž., Banovac, A., Kružić, I., & Jerković, I. (2023). ChatGPT-3.5 as writing assistance in students' essays. *Humanities and Social Sciences Communications, 10*, 750. https://doi.org/10.1057/s41599-023-02269-7

Braun, V., & Clarke, V. (2021). One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology, 18*(3), 328 352. https://doi.org/10.1080/14780887.2020.1769238

Chan, C. K. Y., & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education, 20*, 43. https://doi.org/10.1186/s41239-023-00411-8

Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International, 61*(2), 228 239. https://doi.org/10.1080/14703297.2023.2190148

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., et al. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management, 71*, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education, 20*, 57. https://doi.org/10.1186/s41239-023-00425-2

Farrokhnia, M., Banihashem, S. K., Noroozi, O., & Wals, A. (2024). A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innovations in Education and Teaching International, 61*(3), 460 474. https://doi.org/10.1080/14703297.2023.2195846

Guo, K., & Wang, D. (2024). To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing. *Education and Information Technologies, 29*(7), 8435 8463. https://doi.org/10.1007/s10639-023-12146-0

Jeon, J., & Lee, S. (2023). Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT. *Education and Information Technologies, 28*(12), 15873 15892. https://doi.org/10.1007/s10639-023-11834-1

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

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

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

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

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics, 2*(2), 100050. https://doi.org/10.1016/j.rmal.2023.100050

Nasaruddin, N. (2024). Using ChatGPT in teaching Arabic as a foreign language. *Arabiyatuna: Jurnal Bahasa Arab, 8*(1), 1 24. https://doi.org/10.29240/jba.v8i1.9413

Ngo, T. T. A. (2023). The perception by university students of the use of ChatGPT in education. *International Journal of Emerging Technologies in Learning, 18*(17), 4 19. https://doi.org/10.3991/ijet.v18i17.39019

Ouyang, F., Zheng, L., & Jiao, P. (2022). Artificial intelligence in online higher education: A systematic review of empirical research from 2011 to 2020. *Education and Information Technologies, 27*(6), 7893 7925. https://doi.org/10.1007/s10639-022-10925-9

Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching, 6*(1), 342 363. https://doi.org/10.37074/jalt.2023.6.1.9

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction, 91*, 101894. https://doi.org/10.1016/j.learninstruc.2024.101894

Strzelecki, A. (2024). To use or not to use ChatGPT in higher education? A study of students' acceptance and use of technology. *Interactive Learning Environments, 32*(9), 5142 5155. https://doi.org/10.1080/10494820.2023.2209881

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

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

Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., & Agyemang, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learning Environments, 10*, 15. https://doi.org/10.1186/s40561-023-00237-x

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

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

Zhang, Z. (V.), & Hyland, K. (2022). Fostering student engagement with feedback: An integrated approach. *Assessing Writing, 51*, 100586. https://doi.org/10.1016/j.asw.2021.100586

Zou, M., & Huang, L. (2023). To use or not to use? Understanding doctoral students' acceptance of ChatGPT in writing through technology acceptance model. *Frontiers in Psychology, 14*, 1259531. https://doi.org/10.3389/fpsyg.2023.1259531