JOURNAL OF MULTIDISCIPLINARY RESEARCH AND INTEGRATED SCIENCES

Volume 4, Issue 1, 2024 | **Article ID:** JMRIS-2024-003 | **Pages:** 23-33 | **Article Type:** Systematic Review
Received: 22-01-2024 | **Revised:** 05-02-2024 | **Accepted:** 15-02-2024 | **Published:** 15-03-2024

Generative AI-Assisted Feedback and Academic Writing Performance Among Undergraduate Students: A Systematic Review of Controlled Comparative Evidence

Emilia John1*,

1 UCL Institute of Education, University College London, London, United Kingdom

***Corresponding author:** Emilia John, UCL Institute of Education, University College London, London, United Kingdom | MJ12415@gmail.com

ABSTRACT

Background: Generative artificial intelligence can provide immediate, scalable, and highly personalized feedback on student writing, but the educational value of such feedback depends on whether students use it to revise and learn rather than simply outsource composition. Evidence emerging in 2023 produced both positive and neutral controlled findings. **Objective:** To synthesize high-quality evidence available through 31 December 2023 concerning the effect of generative AI-assisted feedback on academic writing performance, with particular attention to undergraduate and university language learners and to comparison with human or conventional instruction. **Methods:** A structured systematic evidence synthesis was conducted using Q1 literature in educational technology, psychology, writing assessment, and computer-assisted language learning. Direct 2023 controlled studies of ChatGPT or GPT-4 were prioritized. Earlier automated writing evaluation research was incorporated as contextual evidence concerning feedback mechanisms. Because studies differed in intervention design, learner population, outcome instruments, and statistical reporting, no new pooled effect was calculated. **Results:** Direct evidence was mixed. In a randomized three-month study of 50 Chinese bachelor's EFL students, ChatGPT-assisted instruction produced higher post-test writing performance than traditional instruction, with a large effect for overall writing (d=0.76) and large effects for organization (d=0.84) and language use (d=0.88). In contrast, a six-week quasi-experiment involving 48 university English-as-a-new-language students found similar writing gains with GPT-4 and human-tutor feedback; the group-by-time interaction was not statistically significant (F=3.094, p=0.085). A small controlled study of free ChatGPT-3.5 writing assistance in master's students likewise found no significant group effect on essay score (p=0.184). A 2023 meta-analysis of pre-generative-AI automated writing evaluation studies reported a large overall effect on writing quality (Hedges g=0.861), establishing that automated feedback can be educationally effective but not proving an equivalent GenAI effect. **Conclusion:** Generative AI can improve academic writing when embedded in structured instruction, but direct evidence through 2023 does not support an assumption that AI feedback is inherently superior to human feedback. Benefits appear contingent on prompt and rubric design, student proficiency, critical engagement with feedback, iterative revision, and preservation of learner agency. The strongest pedagogical model is GenAI as a feedback scaffold rather than a substitute author.

Keywords: generative artificial intelligence; ChatGPT; academic writing; feedback; undergraduate students; automated writing evaluation; EFL; higher education

1. Introduction

Academic writing is both a means of communication and a mechanism through which university students develop disciplinary reasoning. Effective writing requires the coordination of content, organization, evidence, coherence, language control, and revision. Feedback is therefore central to writing development, but high-quality individual feedback is time intensive and may be difficult to provide repeatedly in large university classes.

Automated writing evaluation predates generative artificial intelligence. Earlier systems used rule-based, statistical, or machine-learning methods to identify linguistic features, provide corrective suggestions, and sometimes estimate writing quality. A 2023 meta-analysis of 26 primary automated-writing-evaluation studies involving 2,468 participants found a large positive overall effect on writing quality, with Hedges g=0.861 [1]. This evidence established that computer-generated feedback can contribute to learning, particularly when students revise their work, but conventional automated writing evaluation is narrower than generative AI.

Large language models altered the feedback problem because they can generate natural-language explanations rather than only flag surface features. A student can request clarification, ask for examples, specify a rubric, or ask the system to focus on argument, organization, grammar, or academic style. Barrot described these affordances as potentially useful for second-language writing while simultaneously highlighting reliability, overdependence, authorship, and academic-integrity concerns

[2]. Su et al. similarly argued that collaborative use of ChatGPT can support argumentative-writing processes when the learner remains actively engaged in planning, evaluating, and revising [3].

Direct empirical evidence began emerging in 2023. Escalante et al. compared GPT-4-generated feedback with human-tutor feedback over six weeks and found no significant difference in learning outcomes [4]. Song and Song reported substantially larger gains among bachelor's EFL students receiving a broader ChatGPT-assisted writing intervention than among students receiving conventional instruction [5]. Bašić et al., however, found no improvement in essay quality or writing speed when students were simply permitted to use ChatGPT-3.5 as a writing assistant [6]. These apparently divergent results suggest that the educational effect may depend less on access to the model itself and more on how the model is integrated into the writing process.

The present review therefore examined controlled and comparative evidence available through 2023. The primary objective was to determine whether generative AI-assisted feedback improves academic writing performance among university learners and, where possible, undergraduate students. A second objective was to identify the instructional conditions under which AI feedback is most likely to support learning without eroding student agency or replacing the cognitive work that writing is intended to develop.

2. Methods

2.1 Review design

A structured systematic evidence synthesis was undertaken using literature published no later than 31 December 2023. The review emphasized controlled comparative studies of generative AI in university writing and supplemented them with high-quality automated writing evaluation and writing-feedback research. The focus was educational performance rather than detection of AI-generated text.

2.2 Eligibility criteria

Studies were eligible when they examined university or post-secondary learners, AI-assisted writing or automated feedback, and outcomes directly related to writing quality, revision, writing motivation, or engagement with feedback. Direct ChatGPT or GPT-4 studies were prioritized. Earlier automated writing evaluation studies were used to contextualize mechanisms rather than to estimate the effect of generative AI itself. Studies outside Q1 journals were not used as core evidence.

The evidence eligibility framework is summarised in Table 1.

2.3 Outcome interpretation

Writing improvement was interpreted according to the outcome measure used by each study. Standardized effects were reported only where published by the source study. A de novo meta-analysis was not undertaken because the direct GenAI studies differed substantially in intervention intensity, assignment structure, feedback source, duration, student level, and scoring method. In particular, broad AI-assisted instruction was not treated as equivalent to feedback-only use.

3. Results

3.1 Controlled comparative evidence

The direct evidence available through 2023 was small but informative. Song and Song enrolled 50 Chinese EFL students in bachelor's programs, matched them for proficiency, and randomly assigned them to AI-assisted or conventional instruction for approximately three months [5]. The experimental condition integrated ChatGPT into academic writing practice, including real-time feedback on grammar, vocabulary, sentence structure, coherence, and organization. The control group received traditional instruction from the same teacher using the same instructional materials.

Post-test overall writing scores were 59.12 (SD 14.23) in the experimental group and 45.18 (SD 15.62) in the control group. The reported between-group effect for overall writing was large ($d=0.76$, $p<0.001$). Effects were also substantial for content ($d=0.65$), organization ($d=0.84$), and language use ($d=0.88$), while writing motivation showed a medium effect ($d=0.52$). These findings provide the strongest direct undergraduate evidence of benefit available within the 2023 cutoff. However, the intervention was broader than automated feedback alone and included structured AI-assisted learning, practice, and instructional integration.

Escalante et al. examined a narrower feedback question [4]. Forty-eight university English-as-a-new-language students participated in a six-week quasi-experimental study. The experimental group received GPT-4-generated feedback and the control group received human-tutor feedback on recurring academic writing. Both groups improved, with mean score gains of 5.848 in the AI group and 6.86 in the human-feedback group. The group-by-time interaction was not statistically significant ($F=3.094$, $p=0.085$, partial eta-squared=0.063), and the between-groups effect was also nonsignificant ($F=0.241$, $p=0.626$). Thus, GPT-4 feedback appeared educationally viable but not superior to human feedback.

Bašić et al. provide an instructive contrast [6]. Eighteen second-year master's students completed an essay task, with nine permitted to use ChatGPT-3.5 freely and nine writing conventionally. Mean essay scores were 2.00 in the ChatGPT group and 2.39 in the control group, and writing duration was similar. The group effect was not significant ($p=0.184$). Because the study evaluated general writing assistance rather than feedback-only use and involved postgraduate students, it is best interpreted as contextual evidence that unrestricted access to a generative model does not automatically improve writing.

Direct and contextual comparative studies relevant to generative AI-assisted writing are summarised in Table 2.

Key quantitative results from the principal 2023 controlled university studies are summarised in Table 3.

Writing-score gains with GPT-4 versus human-tutor feedback in the 2023 Escalante et al. university study are presented in Figure 1.

3.2 Student perceptions and engagement with feedback

Student acceptance was generally favorable but not uniformly pro-AI. In Escalante et al.'s separate preference study, the six-week average number preferring human feedback was 18 compared with 19.667 preferring AI feedback [4]. Mean ratings for satisfaction, clarity, helpfulness, and preference were all above approximately 4 on a five-point scale for both sources and differed little. Students valuing human feedback emphasized interaction and the ability to ask follow-up questions, while students favoring AI emphasized clarity, consistency, specificity, academic style, and vocabulary.

These findings are consistent with earlier writing-feedback research showing that the educational effect of feedback depends on student engagement rather than simple receipt. Zhang and Hyland demonstrated that students engage with automated and teacher feedback affectively, behaviorally, and cognitively, and that stronger engagement can be associated with better writing development [9]. Automated feedback is therefore best conceptualized as information that must be interpreted and acted upon, not as an intervention that mechanically improves text.

3.3 What generative AI adds to conventional automated writing evaluation

The Zhai and Ma meta-analysis demonstrates that automated feedback systems were already capable of producing meaningful writing gains before widespread generative AI [1]. Generative models extend

this capability by producing natural-language explanations, responding to follow-up questions, adapting tone and complexity, and addressing higher-level features such as argument and organization. At the same time, these capabilities create new risks because the system can generate plausible but inaccurate advice rather than merely apply a predefined correction rule.

Barrot identified the principal educational tension: generative AI can provide immediate and personalized assistance, but learners may over-rely on it, accept inaccurate feedback, or allow model-generated text to dilute authorial voice [2]. Su et al. similarly framed ChatGPT as a collaborator whose value depends on how the learner participates in argumentative-writing decisions [3]. The central educational variable is therefore not the quantity of AI output but the degree of learner evaluation and revision.

Educational affordances and risks of generative AI-assisted writing feedback are summarised in Table 4.

A blended human–GenAI feedback framework for undergraduate academic writing is presented in Figure 2.

3.4 Recommended controlled-trial architecture

The mixed direct evidence indicates that future studies should isolate the feedback component more carefully. A strong trial would randomize undergraduate students to human feedback, GenAI feedback, blended human-plus-GenAI feedback, and a conventional minimal-feedback control. All groups should receive the same assignments, rubric, teaching time, and revision opportunities. The generative-AI condition should use a fixed, archived prompt and model version, while each generated response should be retained for audit.

The primary endpoint should be writing performance on an independently administered post-test for which generative AI is not available. This design separates learning from assisted task performance. If students are assessed only on the same essays that were edited with AI, higher scores may reflect better editing rather than stronger transferable writing competence. Delayed follow-up should also evaluate whether gains persist after AI support is removed.

The recommended design for a definitive undergraduate controlled trial is presented in Table 5.

4. Discussion

4.1 Principal findings

The direct evidence available through 2023 does not support a simple conclusion that generative AI feedback either improves or fails to improve academic writing. Instead, the findings indicate that intervention design determines whether AI becomes an educational scaffold or merely another source of text. Song and Song demonstrated substantial gains when ChatGPT was embedded within a structured, sustained instructional environment [5]. Escalante et al. found that carefully prompted GPT-4 feedback produced learning outcomes broadly comparable with human tutoring rather than superior outcomes [4]. Bašić et al. found no benefit when students simply had access to ChatGPT during an essay task [6].

These results are not contradictory. They describe different pedagogical uses. Structured AI-assisted instruction creates repeated opportunities for feedback, practice, and revision. Feedback-only AI may substitute for some tutoring functions without producing a large additional advantage. Unstructured writing assistance may reduce neither cognitive load nor task time and can even distract inexperienced users. The key variable is therefore the student's cognitive engagement with the feedback.

4.2 Feedback should improve the writer, not only the text

A central methodological issue is whether an intervention improves the submitted document or the learner's underlying writing competence. Generative AI can produce sentence-level improvements immediately. If a student simply accepts replacements, the final essay may improve without the student acquiring transferable skill. Educational research must therefore assess performance after the assistance is removed.

This distinction also explains why student agency matters. The strongest learning pathway requires the student to interpret the feedback, compare it with the rubric, decide whether it is valid, revise intentionally, and articulate why the revision improved the text. This process converts AI output into metacognitive practice. By contrast, replacing a paragraph with model-generated prose minimizes precisely the reasoning that writing instruction is intended to cultivate.

4.3 Human feedback retains distinctive advantages

The near-even preference split observed by Escalante et al. suggests that students perceive genuine strengths in both human and AI feedback [4]. Human tutors can interpret assignment history, negotiate meaning in real time, respond to affect, and recognize a student's developing voice. Generative AI can provide consistency, immediate availability, detailed explanations, and repeated reformulation without increasing instructor workload. These characteristics support a blended model rather than a universal replacement model.

Earlier automated feedback research supports the same conclusion. Link et al. examined automated writing evaluation alongside teacher feedback and revision, illustrating that automated systems function within a broader ecology of instruction rather than independently of it [10]. Chen and Cui's comparison of automated and peer feedback likewise showed that different feedback sources may influence different dimensions of cohesion and coherence [11]. Source diversity may therefore be educationally advantageous when students are taught how to reconcile conflicting suggestions.

4.4 Risks of over-reliance and feedback error

Generative AI differs from conventional feedback systems because it can confidently generate novel explanations. This flexibility is pedagogically attractive but introduces uncertainty. A student with limited subject or language knowledge may be unable to identify a plausible but inappropriate recommendation. Kasneci et al. emphasized that large language models in education create opportunities for personalized support while also raising concerns around reliability, bias, assessment, and dependency [12].

For academic writing, feedback accuracy should be judged at multiple levels. Surface corrections may be easy to verify, whereas advice about evidence, argument, disciplinary conventions, and citation can be more difficult. Higher-order feedback should therefore be treated as advisory and subjected to student or instructor verification. Prompt design can reduce but not eliminate this limitation.

4.5 Implications for undergraduate teaching

The evidence favors embedding GenAI within an explicit feedback protocol. Students should first produce an independent draft, then request feedback using assignment-specific criteria, classify the feedback by issue type, evaluate which suggestions are valid, revise selectively, and document important decisions. Instructors can assess not only the final essay but also the quality of the student's revision reasoning. Such a structure preserves the educational purpose of writing while taking advantage of rapid automated feedback.

The approach is particularly promising for EFL students because immediate explanations can reduce barriers associated with grammar, vocabulary, and academic register. However, the larger effects reported in EFL contexts should not automatically be generalized to native-English-speaking students or discipline-specific expert writing. Proficiency, genre, and disciplinary convention are likely moderators and require direct study.

4.6 Strengths and limitations

This synthesis uses a strict pre-2024 evidence window and separates direct generative-AI trials from older automated-writing-evaluation evidence. It also distinguishes feedback-only interventions from broader AI-assisted instruction and unrestricted writing assistance, preventing non-equivalent uses from being treated as a single intervention.

The main limitation is the small direct evidence base available by the end of 2023. Sample sizes were modest, intervention durations were short, and populations were concentrated in language-learning contexts. The Song and Song intervention included more than feedback alone, while the Bašić study involved master's students and a very small sample. Model versions and prompt designs also changed rapidly, creating reproducibility challenges. A formal meta-analysis of the direct GenAI studies would therefore produce misleading precision.

4.7 Conclusion

Generative AI-assisted feedback has substantial potential to support academic writing among undergraduate students, but its educational effect depends on how it is used. The strongest 2023 undergraduate controlled evidence showed meaningful gains when ChatGPT was embedded within sustained AI-assisted instruction, whereas feedback-only GPT-4 produced outcomes comparable to human tutoring and unstructured ChatGPT access did not improve essay quality.

The evidence therefore supports GenAI as a scaffold for revision rather than a substitute author. Effective implementation should preserve an independent first draft, align prompts with explicit rubrics, require students to evaluate rather than automatically accept suggestions, incorporate human oversight for higher-order issues, and assess transfer on writing tasks completed without AI assistance. Future controlled trials should determine whether blended human-plus-GenAI feedback can combine the scalability and specificity of automated feedback with the contextual judgment, dialogue, and pedagogical sensitivity of human instruction.

Declarations

Funding

No external funding was received.

Conflict of Interest

The authors declare no conflict of interest.

Ethical Approval

Not applicable. This manuscript synthesizes previously published evidence and involves no direct participant recruitment.

Consent to Participate

Not applicable.

Data Availability

No novel participant-level dataset was generated or analyzed. All quantitative findings were obtained from previously published studies.

Author Contributions

The authors contributed to conceptualization, evidence synthesis, methodological interpretation, manuscript preparation, and critical revision.

References

1. Zhai N, Ma X. The effectiveness of automated writing evaluation on writing quality: a meta-analysis. J Educ Comput Res. 2023;61(4):875-900. doi:10.1177/07356331221127300.

2. Barrot JS. Using ChatGPT for second language writing: pitfalls and potentials. Assess Writ. 2023;57:100745. doi:10.1016/j.asw.2023.100745.

3. Su Y, Lin Y, Lai C. Collaborating with ChatGPT in argumentative writing classrooms. Assess Writ. 2023;57:100752. doi:10.1016/j.asw.2023.100752.

4. Escalante J, Pack A, Barrett A. AI-generated feedback on writing: insights into efficacy and ENL student preference. Int J Educ Technol High Educ. 2023;20:57. doi:10.1186/s41239-023-00425-2.

5. Song C, Song Y. Enhancing academic writing skills and motivation: assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students. Front Psychol. 2023;14:1260843. doi:10.3389/fpsyg.2023.1260843.

6. Bašić Ž, Banovac A, Kružić I, Jerković I. ChatGPT-3.5 as writing assistance in students' essays. Humanit Soc Sci Commun. 2023;10:750. doi:10.1057/s41599-023-02269-7.

7. Gayed JM, Carlon MKJ, Oriola AM, Cross JS. Exploring an AI-based writing assistant's impact on English language learners. Comput Educ Artif Intell. 2022;3:100055. doi:10.1016/j.caeai.2022.100055.

8. Hwang WY, Nurtantyana R, Purba SWD, Hariyanti U, Indrihapsari Y, Surjono HD. AI and recognition technologies to facilitate English as foreign language writing for supporting personalization and contextualization in authentic contexts. J Educ Comput Res. 2023;61(5):1008-1035. doi:10.1177/07356331221137253.

9. Zhang ZV, Hyland K. Student engagement with teacher and automated feedback on L2 writing. Assess Writ. 2018;36:90-102. doi:10.1016/j.asw.2018.02.004.

10. Link S, Mehrzad M, Rahimi M. Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement. Comput Assist Lang Learn. 2022;35(4):605-634. doi:10.1080/09588221.2020.1743323.

11. Chen M, Cui Y. The effects of AWE and peer feedback on cohesion and coherence in continuation writing. J Second Lang Writ. 2022;57:100915. doi:10.1016/j.jslw.2022.100915.

12. Kasneci E, Sessler K, Küchemann S, Bannert M, Dementieva D, Fischer F, et al. ChatGPT for good? On opportunities and challenges of large language models for education. Learn Individ Differ. 2023;103:102274. doi:10.1016/j.lindif.2023.102274.

Table 1. Evidence eligibility framework

Domain	Inclusion criteria	Exclusion criteria
Publication date	Published on or before 31 December 2023	2024 onward
Journal quality	Q1 journal in education, writing, psychology, AI, or language learning	Lower-quartile or unranked journal
Population	Undergraduate or university students; postgraduate evidence only as contextual comparison	School-only populations unless included in a high-quality meta-analysis
Intervention	ChatGPT/GPT feedback, AI writing assistance, automated writing evaluation	AI use unrelated to writing or feedback
Comparator	Human feedback, conventional teaching, non-AI instruction, peer feedback, or pre/post comparison	No interpretable educational comparator
Outcomes	Writing quality, content, organization, language use, revision, motivation, feedback engagement	Perception alone without learning relevance
Synthesis	Controlled evidence prioritized; heterogeneous designs synthesized descriptively	Artificial pooling of non-equivalent outcomes

Table 2. Direct and contextual comparative studies relevant to generative AI-assisted writing

Study	Population/design	AI condition	Comparator	Principal result
Song & Song [5]	50 bachelor's EFL students; randomized; 3 months	ChatGPT-assisted writing instruction with real-time feedback	Traditional instruction	Overall writing d=0.76; organization d=0.84; language use d=0.88; all favored AI condition
Escalante et al. [4]	48 university ENL students; 6-week quasi-experiment	GPT-4-generated writing feedback	Human-tutor feedback	Both improved; group × time F=3.094, p=0.085; no significant superiority
Bašić et al. [6]	18 second-year master's students; controlled essay task	Free ChatGPT-3.5 writing assistance	Conventional writing	No significant group effect on essay score, p=0.184
Zhai & Ma [1]	26 AWE studies; 2,468 participants	Automated writing evaluation feedback	Varied controls	Large pooled effect on writing quality, g=0.861; contextual predecessor rather than GenAI estimate
Gayed et al. [7]	English language learners; AI writing-assistant intervention	AI-supported writing environment	Conventional/alternative condition	AI writing support improved performance-related outcomes and reduced barriers to writing
Hwang et al. [8]	EFL learners in authentic contexts	AI and recognition technologies supporting personalized writing	Comparison condition	Technology-supported personalization improved relevant writing-learning processes

Table 3. Key quantitative results from the two principal 2023 controlled university studies

Outcome	AI condition	Comparator	Published test/effect	Interpretation
Overall writing, Song & Song [5]	59.12 (SD 14.23)	45.18 (SD 15.62)	d=0.76; p<0.001	Large advantage for AI-assisted instruction
Content, Song & Song [5]	15.96 (SD 3.71)	13.71 (SD 3.12)	d=0.65; p=0.003	Medium-to-large advantage
Organization, Song & Song [5]	16.56 (SD 3.54)	13.63 (SD 3.63)	d=0.84; p<0.001	Large advantage
Language use, Song & Song [5]	19.89 (SD 4.82)	15.89 (SD 4.12)	d=0.88; p<0.001	Large advantage
Writing motivation, Song & Song [5]	20.06 (SD 3.33)	18.21 (SD 3.58)	d=0.52; p=0.001	Medium advantage
Writing gain, Escalante et al. [4]	+5.848	+6.86	Group × time p=0.085	Comparable improvement with AI and human feedback
Feedback source, Escalante et al. [4]	GPT-4	Human tutor	Between-group p=0.626	No significant source effect

Table 4. Educational affordances and risks of generative AI-assisted writing feedback

Dimension	Potential educational value	Primary risk	Design response
Immediacy	Feedback is available during drafting and revision	Students may revise before reflecting independently	Require an initial independent draft
Specificity	Prompts can target organization, grammar, argument, or style	Overly narrow prompts may miss global writing problems	Align prompts with assignment rubric
Dialogue	Students can request clarification or examples	Repeated prompting can drift into answer generation	Use questions that explain rather than rewrite
Personalization	Feedback can be adjusted to proficiency and needs	Inaccurate personalization may create false confidence	Require student verification and instructor oversight
Scalability	Large classes can receive rapid formative feedback	Volume may be mistaken for quality	Evaluate accuracy and usefulness, not response length
Language support	Can reduce linguistic barriers for EFL learners	May homogenize voice or produce unnatural register	Preserve student wording where possible
Higher-order feedback	Can comment on coherence and argumentation	Model may invent reasoning or misread disciplinary context	Keep student responsible for content decisions
Revision support	Can explain why a change may improve text	Copying model revisions can reduce cognitive engagement	Grade rationale and revision process as well as final text

Table 5. Recommended design for a definitive undergraduate controlled trial

Component	Recommended design	Rationale
Population	Undergraduate students stratified by baseline writing proficiency	Allows subgroup analysis by learner level
Study arms	Human feedback; GenAI feedback; blended feedback; minimal-feedback control	Separates AI effect from feedback exposure itself
Intervention	Same writing tasks, rubric, feedback frequency, and revision window	Reduces instructional confounding
AI specification	Archived model version, fixed system prompt, documented user prompt	Improves reproducibility
Primary outcome	Blind-rated independent post-test writing without AI access	Measures transferable competence rather than edited product quality
Secondary outcomes	Content, organization, language use, revision quality, motivation, self-efficacy	Identifies mechanisms of benefit
Process measures	Prompt logs, feedback uptake, accepted/rejected changes, revision rationale	Measures cognitive engagement
Integrity measures	Text provenance and student explanation of revisions	Distinguishes assistance from substitution
Follow-up	Delayed writing test after 4–8 weeks	Measures retention
Analysis	ANCOVA or mixed-effects model controlling baseline performance; preregistered subgroup analyses	Provides stronger causal inference

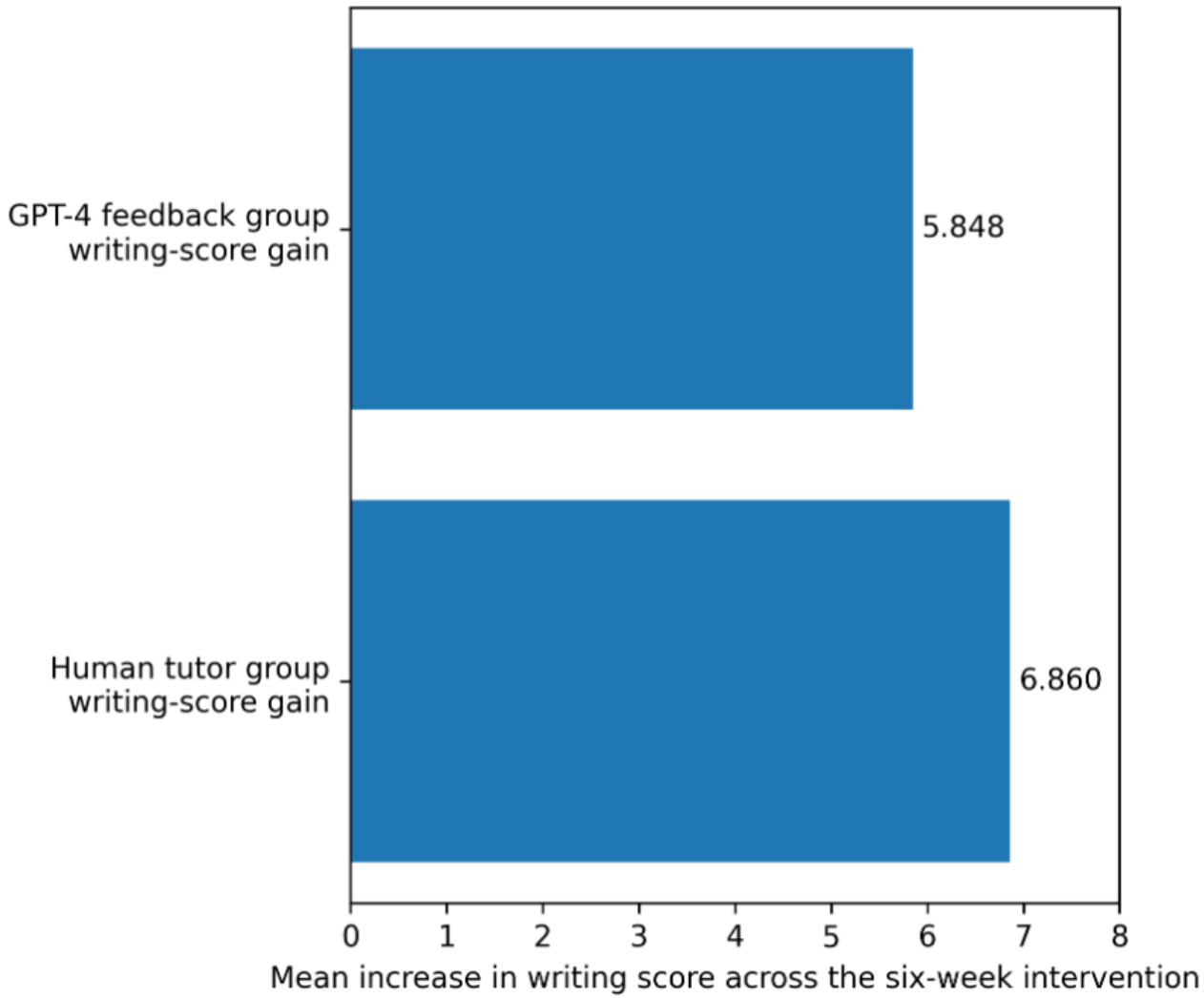
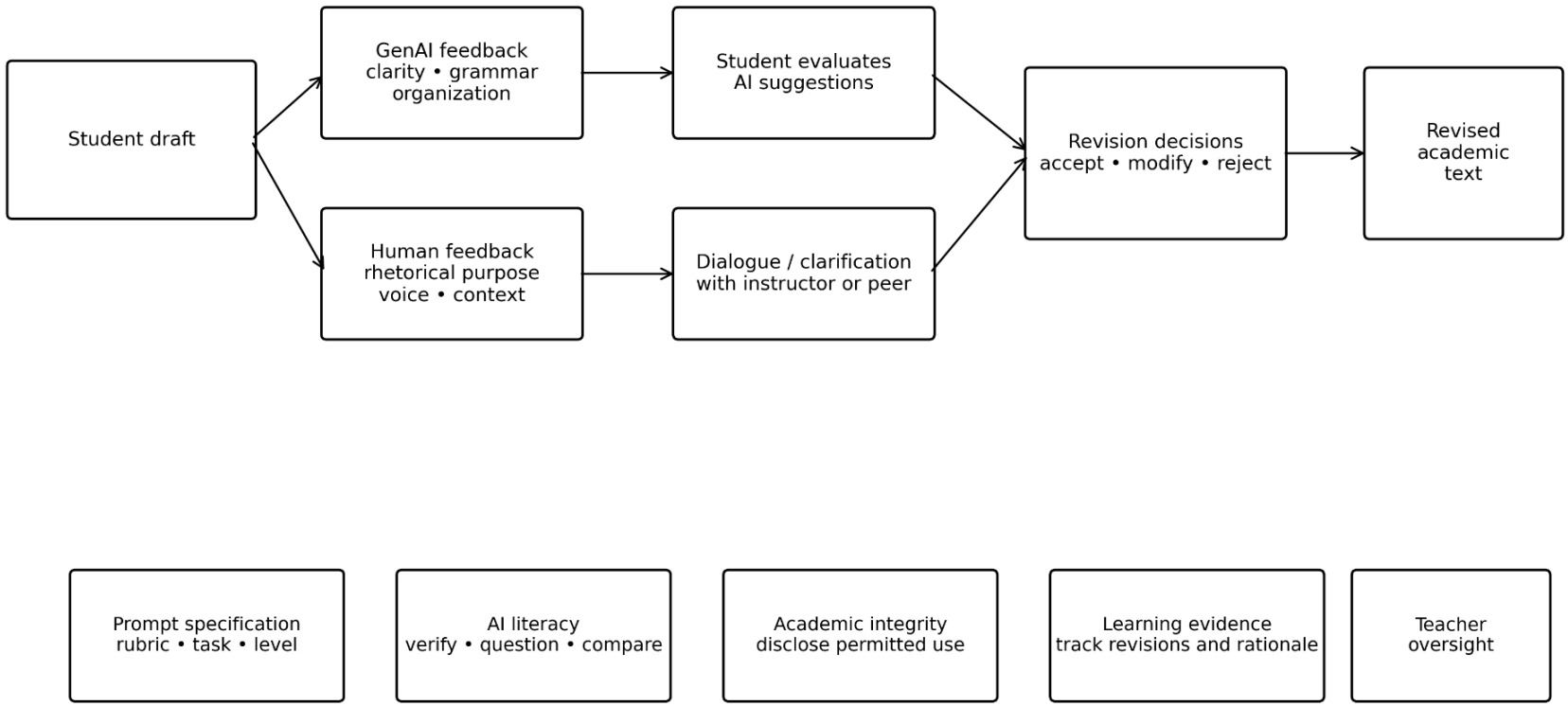


Figure 1. Writing improvement with GPT-4 versus human-tutor feedback in a 2023 university study.

Escalante et al. reported mean writing-score gains of 5.848 in the GPT-4 feedback group and 6.860 in the human-tutor group. Both groups improved, while the group-by-time interaction (F=3.094, p=0.085) and between-group effect (F=0.241, p=0.626) were not statistically significant.

Figure 2. Blended human–GenAI feedback framework for undergraduate academic writing



The pedagogical objective is not automatic rewriting. Effective AI-assisted feedback keeps the student responsible for evaluating feedback and making revision decisions.

Figure 2. Blended human–GenAI feedback framework for undergraduate academic writing.

The framework preserves learner agency by requiring students to evaluate AI and human feedback before accepting, modifying, or rejecting suggested revisions. Prompt specification, AI literacy, academic-integrity safeguards, learning evidence, and teacher oversight act as implementation controls.

How to cite: John E. (2024). Generative AI-Assisted Feedback and Academic Writing Performance Among Undergraduate Students: A Systematic Review of Controlled Comparative Evidence. Journal of Multidisciplinary Research and Integrated Sciences, 4(1).