



Generative Artificial Intelligence in Academic Writing Instruction: What Classroom Studies Show So Far

Hai Nikolic^{1*}, Jakub Teixeira², Hala Ristic³

¹Department of Social Sciences, Dalmatian Institute for Advanced Social Sciences, Brazil

²Department of Social Sciences, Danube Institute for Advanced Social Sciences, Oman

³Department of Social Sciences, Niger Academy of Social Sciences, Denmark

Article Info

Volume 1, Issue 1, 2026

doi: [10.62715/ijetl.2026.v1.i1.pp7-12](https://doi.org/10.62715/ijetl.2026.v1.i1.pp7-12)

Received: 09 February 2026

Accepted: 23 April 2026

Published: 21 June 2026

Abstract

Large language models that produce fluent prose on request became freely available to students at the end of 2022, and writing teachers have been adjusting ever since. Commentary on the change is abundant. Evidence from classrooms is newer and smaller, and it is the subject of this narrative review. The article considers published studies in which generative artificial intelligence was used in, or evaluated for, the teaching of academic writing, mainly in higher education and in English as an additional language. It groups the findings under four uses: support for planning, support for drafting, support for revision, and the provision of feedback and scores. The clearest results concern feedback, where machine comments on student drafts have been found to approach, without matching, the quality of comments by trained human readers. Results on learning are sparser and depend on short interventions and self-report. Studies of detection software give consistent grounds for caution, including a documented bias against writers of English as an additional language. The review sets these findings against what is known about how writing is learned, identifies the weaknesses of the evidence base, and proposes questions that classroom research should now address.

Keywords: *Generative artificial intelligence, Academic writing, Writing instruction, Automated feedback, Large language models, Academic integrity*

© 2026 Hai Nikolic et al. This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

1. Introduction

The public release of ChatGPT in November 2022 put a capable text generator in the hands of anyone with an internet connection. For most school and university subjects this was a change in the tools available. For the teaching of writing it reached the object of instruction itself, because the essay, the report and the literature review are both what students are taught to produce and the means by which their learning is judged. Early responses in the scholarly press were position papers that listed opportunities and threats (Kasneci et al., 2023) or concentrated on cheating (Cotton et al., 2024). A rapid review of the first months of publication dealt mostly with the tool's performance on tasks in different subjects and with its potential uses and risks (Lo, 2023).

* Corresponding author: Hai Nikolic, Department of Social Sciences, Dalmatian Institute for Advanced Social Sciences, Brazil. E-mail: hai.nikolic@dalmatian.edu

© 2026 Hai Nikolic et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Studies conducted with students and teachers have since begun to appear. They are small, they are concentrated in higher education and in second language writing, and they vary a good deal in rigour. They are nonetheless a different kind of source from commentary, and teachers who must decide what to permit, what to teach and what to assess are entitled to know what they show. This article reviews that body of work. Its title promises what classroom studies show so far, and the qualification is meant seriously: the tools change faster than studies can be published, and several conclusions below may not hold for later systems.

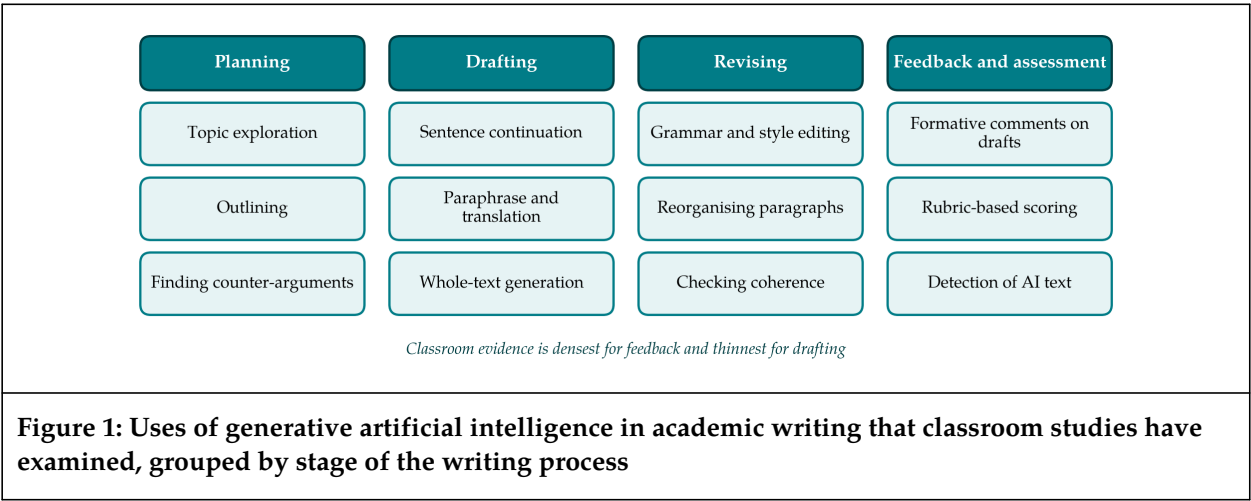
2. Scope and Approach

The review is narrative. It covers peer-reviewed studies and widely cited reports, published since the arrival of general-purpose chat models, that examine generative artificial intelligence in the teaching, learning or assessment of academic writing. Studies of the software alone are included only where they bear directly on a classroom decision, as with scoring and detection. Earlier research on automated writing evaluation and on the psychology of composition is drawn on for context. No systematic search was carried out and no counts of sources are offered; the selection reflects the studies that have shaped discussion in the field, and other reviewers might reasonably weigh them differently.

3. What Writing Instruction Is For

Any judgement about a writing tool rests on a view of what learning to write involves. The cognitive process model of Flower and Hayes (1981) described composing as the recursive interplay of planning, translating ideas into text and reviewing, all under the writer's own monitoring. Bereiter and Scardamalia (1987) distinguished novices, who tell what they know in the order it occurs to them, from experts, who transform their knowledge in the effort to meet a rhetorical goal. On this account the difficulty of writing is where the learning happens. The meta-analysis of instruction for adolescents by Graham and Perin (2007) found the largest gains from teaching strategies for planning, revising and summarising, which are precisely the operations a language model can now perform on a student's behalf.

Two concepts frame the risk and the promise. Cognitive offloading, the use of external aids to reduce mental demand, is efficient for the task in hand and can leave the underlying capacity undeveloped (Risko and Gilbert, 2016). Scaffolding, a term from the sociocultural tradition in educational psychology, is help that lets a learner do with assistance what will later be done alone. The same prompt to a chatbot can be either, and the difference lies in what the student does next. Much of the classroom evidence can be read as an attempt to find out which of the two is occurring. Figure 1 sets out the uses that studies have examined across the stages of composing.



4. Evidence by Use

The studies are grouped below by the part of the writing process they address. Table 1 gives an overview

of what each group reports and how far the evidence extends.

Table 1: Uses of generative artificial intelligence in writing instruction and the state of the classroom evidence			
Use	What studies report	Main limit of the evidence	Example source
Planning	Students value help with ideas and outlines	Perception data; little on text quality	Su et al., 2023
Drafting	Faster production and more fluent text	Learning rarely separated from performance	Noy and Zhang, 2023
Revising	Surface accuracy improves readily	Short interventions, no delayed measures	Yan, 2023
Formative feedback	Machine comments near human quality, less accurate	Uptake by students seldom followed	Steiss et al., 2024
Scoring	Moderate agreement with human raters	Tested on corpora, not in live courses	Mizumoto and Eguchi, 2023
Detection	Unreliable, with bias against non-native writers	Tools and models change quickly	Liang et al., 2023

4.1. Planning and drafting

Work on planning is mostly exploratory. [Su et al. \(2023\)](#) examined how ChatGPT could assist the stages of argumentative writing, from outlining to editing and reflection, and presented the tool as a collaborator whose suggestions require evaluation by the writer. [Kohnke et al. \(2023\)](#) catalogued uses for language teaching and stressed the digital competence both teachers and learners need in order to use the tool critically. [Barrot \(2023\)](#) set out the attractions for second language writers, among them instant responses and adjustable difficulty, together with the pitfalls of formulaic organisation, invented sources and dependence. These papers define the territory. They do not establish what students learn.

On drafting, the most rigorous evidence comes from outside education. In a randomised experiment with professionals carrying out occupational writing tasks, [Noy and Zhang \(2023\)](#) found that access to ChatGPT shortened the time taken and raised rated quality, with the weakest writers gaining most. The result is often cited in educational debate and should be handled carefully there. It concerns the quality of a product made with the tool, by adults who could already write. A classroom is interested in what the writer can do afterwards without it, and the experiment was not designed to say.

4.2. Revision and language support

Students writing in an additional language have been the most studied group, and for an obvious reason: a model that corrects grammar and suggests idiomatic phrasing answers a need they feel keenly. In an exploratory study of a second language writing practicum, [Yan \(2023\)](#) recorded students' recognition of the tool's usefulness alongside their unease about fairness and about what its use would do to the honesty of assessment. [Warschauer et al. \(2023\)](#) named the tensions directly: the tool rewards those who already write well enough to judge and direct its output, and students are penalised for using it in class while being expected to use it at work. Their proposed remedy is a teaching sequence in which learners understand the tool, gain access to it, prompt it, corroborate what it says and incorporate the result with acknowledgement.

Surveys of students give a consistent picture of attitudes. In the Hong Kong survey by [Chan and Hu \(2023\)](#), university students were broadly positive about generative tools for brainstorming and writing support and were at the same time worried about accuracy, privacy and the effect on their own development. Attitudes are worth knowing, but they are a weak proxy for learning. Students are poor judges of how much a fluent aid has taught them.

4.3. Feedback and scoring

Feedback is where the evidence is strongest, partly because the outcome is easier to define. [Steiss et al. \(2024\)](#) compared formative comments written by trained human evaluators with comments generated by ChatGPT on the same secondary school essays. The human comments were rated higher on most criteria, including accuracy and the prioritising of what mattered most, but the differences were modest, and the authors concluded that machine feedback has value where timely human feedback is not available. [Dai et al. \(2023\)](#) found the model's feedback on student project proposals more detailed and readable than the instructor's and in fair agreement with the instructor on what needed attention.

Two studies looked at feedback inside courses. [Escalante et al. \(2023\)](#) compared the writing development of university learners of English who received feedback from a human tutor with that of learners who received it from a language model. Outcomes did not differ, and the students were divided over which source they preferred. [Guo and Wang \(2024\)](#) set ChatGPT's comments on argumentative essays beside those of teachers of English as a foreign language. The model produced more feedback and spread it across content, organisation and language, and the teachers saw it as a useful supplement to their own. These results continue an older line of work on automated writing evaluation, in which [Stevenson and Phakiti \(2014\)](#) found modest evidence that computer-generated feedback improves the quality of revised texts and little evidence of transfer to new writing. The long-standing finding that feedback helps only when students engage with it ([Hyland and Hyland, 2006](#)) applies with equal force to machine comments.

For scoring, [Mizumoto and Eguchi \(2023\)](#) applied a language model to a large corpus of essays by learners of English and obtained scores that agreed with human ratings to a useful but imperfect degree, improved when conventional linguistic measures were added. This supports low-stakes uses, such as giving students an indicative band on a practice essay. It does not support replacing human judgement where grades have consequences.

4.4. Integrity and detection

Many institutions first responded to generative tools by turning to detection software, and the research on detectors is unusually consistent. [Weber-Wulff et al. \(2023\)](#) tested a range of detection tools and judged them neither accurate nor reliable, with a tendency to pass machine text as human and to fail once text had been paraphrased or translated. [Liang et al. \(2023\)](#) showed that detectors misclassified essays by non-native writers of English as machine-generated far more often than essays by native writers, because limited lexical variety resembles the statistical signature the detectors seek. A penalty imposed on the strength of a detector score therefore falls unequally. Guidance from [UNESCO \(2023\)](#) recommends that institutions invest in human capacity, validate tools before adoption and protect learners' agency. Its emphasis falls on regulation and on the competence of teachers, and detection has a small place in it.

5. Limits of the Evidence

Four weaknesses recur. The first is that performance with the tool is taken as evidence of learning. A better essay written with assistance shows that the assistance works; whether the student writes better unaided is a separate question that needs a later, unassisted task to answer. The second is duration. Interventions typically last a few weeks, which is short against the years over which writing develops. The third is reliance on self-report from volunteers in courses where the tool was permitted, who are unlikely to represent students in general. The fourth is the narrowness of the settings: university students of English dominate, and school pupils, writers in other languages and writers in the disciplines are thinly represented.

A further difficulty is specific to this technology. A study names a model version that may be withdrawn before the article is printed, so exact replication is rarely possible. And the models themselves carry known problems that classroom studies seldom measure: they generate plausible text without reference to truth and reproduce patterns in their training data ([Bender et al., 2021](#)), which in academic writing appears as fabricated citations and confident error.

6. Questions for the Next Round of Studies

The most useful studies would compare forms of use, since comparing use with non-use is rapidly becoming artificial. Whether students who ask a model to critique their own draft develop differently from students who ask it to produce a draft is a question with direct consequences for teaching, and it can be examined with delayed, unassisted writing as the outcome. Studies should record the interaction itself, the prompts and the uses made of the responses, because the label "used ChatGPT" covers behaviours as different as copying and arguing. Research is also needed on assessment designs that remain informative when a tool is available: staged submissions, oral defence of written work and writing done under observation are widely recommended and little evaluated.

Teachers are part of the evidence as well. How instructors read machine feedback, when they trust it and how it changes the time they give to each student are questions that only a few studies have raised, and the answers will decide whether feedback tools ease teaching loads or add a further layer of checking.

7. Conclusion

Classroom research on generative artificial intelligence in writing instruction is at an early stage, and its firmest findings are modest ones. Machine feedback on drafts is good enough to be useful where human feedback is scarce and is not yet its equal. Students welcome the tools and are uneasy about them. Detection software cannot bear the weight institutions have placed on it and treats writers of English as an additional language unfairly. On the central matter, whether students who learn to write with these tools become better or worse writers, the studies so far do not give an answer. Teachers should plan on that footing, and researchers should treat the unassisted writing of students, measured after the tool is taken away, as the outcome that matters most.

Conflict of interest

To be declared by the author(s) before publication.

Funding

To be declared by the author(s) before publication.

References

- 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>](https://doi.org/10.1016/j.asw.2023.100745)
- Bender, E.M., Gebru, T., McMillan-Major, A. and Shmitchell, S. (2021). [On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610-623. <https://doi.org/10.1145/3442188.3445922>](https://doi.org/10.1145/3442188.3445922)
- Bereiter, C. and Scardamalia, M. (1987). [The psychology of written composition](#). Lawrence Erlbaum Associates, Hillsdale, NJ.
- Chan, C.K.Y. and 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>](https://doi.org/10.1186/s41239-023-00411-8)
- Cotton, D.R.E., Cotton, P.A. and 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>](https://doi.org/10.1080/14703297.2023.2190148)
- Dai, W., Lin, J., Jin, H., Li, T., Tsai, Y.-S., Gašević, D. and Chen, G. (2023). [Can large language models provide feedback to students? A case study on ChatGPT. *Proceedings of the 2023 IEEE International Conference on Advanced Learning Technologies \(ICALT\)*, 323-325.](#)
- Escalante, J., Pack, A. and 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>

- Flower, L. and Hayes, J.R. (1981). *A cognitive process theory of writing*. *College Composition and Communication*, 32(4): 365-387. <https://doi.org/10.2307/356600>
- Graham, S. and Perin, D. (2007). *A meta-analysis of writing instruction for adolescent students*. *Journal of Educational Psychology*, 99(3): 445-476. <https://doi.org/10.1037/0022-0663.99.3.445>
- Guo, K. and 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: 8435-8463.
- Hyland, K. and Hyland, F. (2006). *Feedback on second language students' writing*. *Language Teaching*, 39(2): 83-101.
- 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. and Zou, D. (2023). *ChatGPT for language teaching and learning*. *RELC Journal*, 54(2): 537-550. <https://doi.org/10.1177/00336882231162868>
- Liang, W., Yuksekogonul, M., Mao, Y., Wu, E. and Zou, J. (2023). *GPT detectors are biased against non-native English writers*. *Patterns*, 4(7): 100779. <https://doi.org/10.1016/j.patter.2023.100779>
- 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>
- Mizumoto, A. and 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>
- Noy, S. and Zhang, W. (2023). *Experimental evidence on the productivity effects of generative artificial intelligence*. *Science*, 381(6654): 187-192. <https://doi.org/10.1126/science.adh2586>
- Risko, E.F. and Gilbert, S.J. (2016). *Cognitive offloading*. *Trends in Cognitive Sciences*, 20(9): 676-688. <https://doi.org/10.1016/j.tics.2016.07.002>
- Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M. and 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>
- Stevenson, M. and Phakiti, A. (2014). *The effects of computer-generated feedback on the quality of writing*. *Assessing Writing*, 19: 51-65.
- Su, Y., Lin, Y. and Lai, C. (2023). *Collaborating with ChatGPT in argumentative writing classrooms*. *Assessing Writing*, 57: 100752.
- UNESCO (2023). *Guidance for generative AI in education and research*. UNESCO, Paris.
- Warschauer, M., Tseng, W., Yim, S., Webster, T., Jacob, S., Du, Q. and Tate, T. (2023). *The affordances and contradictions of AI-generated text for writers of English as a second or foreign language*. *Journal of Second Language Writing*, 62: 101071. <https://doi.org/10.1016/j.jslw.2023.101071>
- Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P. and Waddington, L. (2023). *Testing of detection tools for AI-generated text*. *International Journal for Educational Integrity*, 19: 26. <https://doi.org/10.1007/s40979-023-00146-z>
- Yan, D. (2023). *Impact of ChatGPT on learners in a L2 writing practicum: An exploratory investigation*. *Education and Information Technologies*, 28: 13943-13967. <https://doi.org/10.1007/s10639-023-11742-4>

Cite this article as: Nikolic H., Teixeira J. and Ristic H. (2026). *Generative Artificial Intelligence in Academic Writing Instruction: What Classroom Studies Show So Far*. *International Journal of Education Teaching and Learning*. 1(1), 7-12. doi: 10.62715/ijetl.2026.v1.i1.pp7-12