International Journal of Research in E-learning
Vol. 12 (1), 2026, pp. 1–21

DOI 10.31261/IJREL.2026.12.1.01

Kamil Lemanek
Maria Curie-Skłodowska University, Poland
https://orcid.org/0000-0002-6260-8973

Aleksander Leczkowski
University of Warsaw, Poland
https://orcid.org/0000-0002-7323-4161

Integrating LLMs and Academic Writing Instruction: Engaged Reading and LLM-Generated Material

Abstract

Integrating Large Language Models (LLMs) into course designs in the higher education context remains an open and quickly growing area of research. This study contributes to the emerging literature by examining how students engage with and evaluate LLM-generated feedback within academic writing instruction. Building on prior work on integration and LLM detection, the study adopts a diachronic design to assess whether students can reliably distinguish between human- and LLM-generated reviews, and how engagement with material may affect perceived classification. Fifty-one undergraduate participants received both human and LLM-generated reviews for essays they authored, completing surveys before and after revising their work. The results indicate that students were initially at random odds in distinguishing human from LLM-generated texts but exhibited statistically significant improvement following engagement. Nearly all students reported positive attitudes toward personalized feedback, irrespective of whether they perceived it to be human or LLM-generated. The findings underscore both the pedagogical potential and evaluative ambiguity of LLM integration in writing instruction.

K e y w o r d s: higher education, writing, detection, course design

The arrival of accessible Large Language Models (LLMs) and related generative AI (GenAI) has brought significant new challenges and opportunities to higher education classrooms (see, e.g., Baidoo-Anu & Owusu Ansah, 2023; Grassini, 2023). The emerging field bringing together these new tools and education is complex, addressing the perspectives of instructors and students relative to a quickly developing set of new and increasingly embedded technologies (see, e.g., Qian, 2025; Phokoye et al., 2025). That complexity is ramified by the social, ethical, and technical contexts associated with these models (see Prunkl et al., 2021; Gabriel et al., 2024).

In the context of the higher education classroom and specifically language and writing courses, there has been significant work done in terms of collecting and reporting data concerning instructor and student attitudes relating to LLMs (see Chan & Hu, 2023; Ortiz-Bonnin & Blahopoulou, 2025; Stöhr et al., 2024). There is also a growing literature on how LLMs may be incorporated into course design to support instructors via chatbots and discussion partners (Jeon & Lee, 2023; Rosas & Rodríguez, 2024; Yang et al., 2022).

These distinct streams of research may be brought together to develop course designs with novel integrations of LLMs, which naturally involve work relating to teaching methodology, instructor and student attitudes concerning LLMs, and the state of the art in LLM modeling and development. One might consider a simple scenario that motivates this sort of integration; suppose an instructor recognizes that students would benefit from more comprehensive individualized feedback for a given exercise that goes beyond what an Automated Writing Evaluation (AWE) tool provides (see Ding & Zou, 2024; Liu, 2024). However, given that providing extended feedback is both time- and effort-intensive, the instructor may determine that providing extended feedback is not a practical possibility, particularly for high-volume courses. This could motivate the instructor to reach for LLMs in pursuit of a solution.

Addressing this sort of situation and developing an adequate integration requires consideration not just of the teaching methodology assumed in the course, or the attitudes of instructors and students, but also a clear picture of the quickly developing abilities of state-of-the-art LLMs and GenAI tools. This latter understanding of LLMs includes the literature on detection, or our ability to discern LLM-generated material from material prepared by humans, which supports, on the one hand, the further refining of models relative to assessment metrics and, on the other, the growing need to be able to identify generated material in various contexts (see, e.g., Ma et al., 2023; Wang et al., 2024; cf. Weber-Wulff et al., 2023).

Arguably, the most significant general development in this area in recent years is the emerging consensus that contemporary models are now effective enough to consistently fool human assessors (see Wu et al., 2025; Dugan et al., 2023). In years prior, human assessment was "the gold standard" (cf. Clark et al., 2021), with humans focusing on features of texts like their coherence, their structure, and

ultimately their perceived naturalness (Howcroft et al., 2020; Novikova et al., 2018; van der Lee et al., 2021). However, these features are now mimicked well enough by LLMs that large-scale studies have shown human assessors to be at random odds in determining whether a given text was LLM-generated in most contexts, including across material like news articles, dating bios, recipes, abstractive summaries, poetry, and short stories (see, e.g., Köbis & Mossink, 2021; Soni & Wade, 2023; Clark et al., 2021). One exception to this is a recent study that found that humans were above random odds when assessing abstracts for scientific articles (Ma et al., 2023), which will be relevant here. The exception aside, the findings of the majority of the studies in the space have motivated a shift away from human assessment as a relevant metric for the effectiveness of various language models to automated systems (see, e.g., Wu et al., 2024; Su & Wu, 2024), which underlie many of our institutional and commercial LLM detection tools.

The arrival of LLMs that are able to produce text that humans cannot distinguish from material actually produced by humans carries with it significant potential impetus for the incorporation of LLM-generated material as parts of academic writing courses. The implication here is simply that if these tools can dynamically generate text that is indiscernible from human-written texts, adjusted to a given context (in this case, to academic writing), then they may be used to produce material up to the standard required for producing complex feedback as part of a writing course. With some guidance from an instructor, these models could produce the sort of detailed and personalized feedback that might otherwise not have been practically feasible for instructors to produce themselves.

Present Study

The aim of this study is to explore this potential integration of academic writing instruction and LLM-generated material through the lens of detection or evaluation – i.e., of whether students can discern LLM-generated feedback from human feedback. Where the majority of studies of human text assessment in this context are synchronic and largely rely on simply reading material and then assessing it (see, e.g., Köbis & Mossink, 2021; Ma et al., 2023; cf. Clark & Smith, 2021), the present study looks to approach the issue diachronically and with more sustained subject engagement, aligning with the sort of treatment LLM-generated content would invite if integrated into an academic writing course.

Importantly, this approach moves beyond the standard emphasis on metrics found in the LLM-detection literature, which is not focused on pedagogical applications but on further developing existing models, with extensions to education being largely incidental (see Wu et al., 2025). In contrast, the present study is

intentionally situated in the target context of an academic writing classroom, and it also incorporates an extended window of assessment and interaction with the material, as noted above.

Moreover, and turning to the literature on education, by drawing on detection as its guiding perspective, the present study contributes a novel approach to work on integrating writing instruction with LLMs, joining recent qualitative analyses and discussions on the subject (Bedington et al., 2024). It also contributes to education research concerning automated feedback, introducing an approach specific to writing instruction, and also providing a side-by-side comparison of human and automated feedback, neither of which appears to have been reported in the literature to this point (cf., e.g., Du et al., 2024; Tzirides et al., 2024). Taken together, the emphasis on teaching, diachronicity, and engagement informs three research questions:

Q1 – *Can students reliably detect LLM-generated material in an academic writing context?*

The first question connects with the established consensus concerning the inability of humans to detect LLM-generated material in short stories, poems, news articles, etc. Assuming that academic writing represents a distinct style or type of writing, this is an important baseline question. It is also potentially substantive to the broader literature on detection, providing further data on human assessment in the context of scientific writing, for which there are findings contrary to the consensus, as noted above (Ma et al., 2023).

Q2 – *Does engaging with the material make a difference in how students assess material?*

The second question builds on the intended diachronicity and engagement with material, aiming to explore whether working with the material at issue makes a difference to how it is perceived – and whether it makes any difference to how they assess its source.

Q3 – *Do students like personalized feedback, and is their impression correlated with their perception of the material as either LLM-generated or not?*

The third question bridges to the student impressions of the introduction of LLM-generated content as something they need to directly interact with, and in the process providing a simple but important practical datapoint for this sort of integration for the future.

Methodology

Participants

Sixty-one undergraduate cognitive science and philosophy students enrolled in English-language academic writing courses conducted at Maria Curie-Skłodowska University and the University of Warsaw participated in the study. Fifty-one completed the requisite surveys and were analyzed (N = 51). All of the students were included in the study by default and explicitly informed that they could opt out at any time. No students chose to opt out.

The final sample comprised 33 women, 15 men, one student identifying as "other," and two who preferred not to answer. Ages ranged from 18 to 35 years (median = 21). Self-reported English proficiency presented as follows: basic (n = 4), intermediate (n = 12), upper-intermediate (n = 16), advanced (n = 18), native (n = 1).

Design and Procedure

The study was conducted as part of two sets of English language academic writing courses. It was embedded in the unit on peer review and publishing in each course as part of an assignment. Each student was required to submit a short essay, following best practices discussed earlier in the course, for which two reviews were prepared. Students were asked to engage with the reviews, making changes to their original manuscripts and preparing separate responses to each review. They were then to submit the revised manuscript and review responses to be graded in light of the course material covered in the unit.

The study followed a within-subjects design. Of the two reviews received by each student, one was authored by the course instructor (human), and one was generated by ChatGPT (GPT-4/XX). The instructor was a native English speaker with experience in teaching academic writing. The participants were given no cues about the authorship of these reviews.

The study consisted of two phases (Survey 1, Survey 2). The first phase took place with the initial distribution of reviews. The participants were asked to read their reviews, with each review being followed by the same questionnaire concerning the material. Following the completion of the questionnaire, students received their reviews as separate files. They were then expected to engage with the reviews, considering their contents, making changes to their manuscripts, and responding to them. The second phase coincided with the submission of the students' revised manuscripts and review responses. The students were presented once again with their original reviews and a field for their responses, which they would already have prepared in advance of submission. Each review/response was again followed by a questionnaire.

To mitigate sequential biases, students were randomly assigned to one of two fixed orders used in both surveys, which determined whether the first review they saw was prepared by the human instructor or LLM-generated: Group 1 (AI, then human) or Group 2 (human, then AI).

Both the instructor and ChatGPT were given prompt guidelines, emphasizing simple but academically appropriate language, suggesting two improvements the student could make, and setting the length of reviews to around 250 words. The ChatGPT prompts also included a set of additional limitations to match standard review style and format, namely, not to use bullet points and not to refer to the author in the second person. One additional limitation included in the prompt was not to suggest adding formal citations/references, as it was out of scope for the assignment in context.

Survey Questionnaires

The contents of the questionnaire for Survey 1 were as follows, reflecting common qualitative criteria considered in human assessment studies (see Howcroft et al., 2020; Novikova et al., 2018):

Consent form, followed by demographic information (gender, age, English level).
First review presented, per the group of the participant.
Three questions on a Likert scale:
Was Review X easy to read and understand? (1 = very difficult, 5 = very easy)
How intuitive was the structure of Review X? (1 = very unintuitive, 5 = very intuitive)
How natural did Review X seem to you? (1 = very unnatural, 5 = very natural)
On the next page, there is a forced-choice source classification question: Which option do you believe is true? Review X was written by a human. // Review X was written by an AI.
The above was repeated for the second review text.
The last question appears on one further page, and it reads as follows: The reviews were personalized for each student – so each student received unique feedback. Did you find that interesting or more engaging? Yes, it was more interesting/engaging for me. // Not really, it didn't make much difference.

The above was repeated for the second review text.

The questionnaire for Survey 2 was slightly different. In addition to those above, it included the following among the qualitative questions:

How easy was it to respond to the review? (1 = It was not easy at all, 5 = It was very easy)

It also featured the forced-choice source classification question on the same page as the qualitative questions. It read as follows: Which option do you believe is true? Review X was written by a human // Review X was written by an AI.

Hypotheses

H1: Survey 1 classification accuracy differs from chance (50%).
H2: Survey 2 classification accuracy differs from chance (50%).
H3: Classification accuracy improves from Survey 1 to Survey 2.

These jointly address Q1 and Q2. Q3 is addressed by the gathered data via the last question of the survey, without the need for a discrete hypothesis.

Analytic Approach

Analyses were conducted in R. For H1 and H2, we tested whether classification (n = 102) accuracy differed from chance using two-sided exact binomial tests with Clopper–Pearson confidence intervals and Cohen's *h* as an effect size.

For H3 (directional hypothesis: Survey 2 > Survey 1), three complementary approaches were used:

H3-A1: Per-guess analysis (n = 102 guesses): Because each student made two classifications per survey, we treated each classification as a trial. Improvements versus declines across surveys were compared using a McNemar test.
H3-A2: Per-student analysis (N = 51 students): For a robustness check, restricted to students who classified both reviews correctly. Changes in "both-correct" status across surveys were again compared using a McNemar test.
H3-B: Per-student analysis (N = 51 students): Each student's accuracy score was computed as the proportion of their two correct classifications in a given survey (0, 0.5, or 1). Differences between Survey 2 and Survey 1 accuracy were tested with a Wilcoxon signed-rank test.

For the Likert-scale outcomes, responses were treated as ordinal within-subject data and analyzed with Wilcoxon signed-rank tests, separately for three tiers: all respondents (N = 51); "both-balanced" (classified one as human, one as AI within a survey; n = 38); "both-correct in both surveys" (n = 16). To control for multiple comparisons, p-values were adjusted using the Benjamini–Hochberg procedure.

"REAL" indicates grouping by the true source (human vs. AI). "PERCEIVED" indicates grouping by the participant's classification (as human vs. as AI), regardless of underlying accuracy.

Results

Classification Accuracy (H1 & H2)

In Survey 1, 47 of 102 classifications were correct (46.1%), which did not differ from chance (p = .488, 95% CI [.36, .56], Cohen's *h* = 0.08). In Survey 2, 59 of 102 classifications were correct (57.8%), also not different from chance (p = .137, 95% CI [.48, .68], Cohen's *h* = 0.16) (Figure 1).

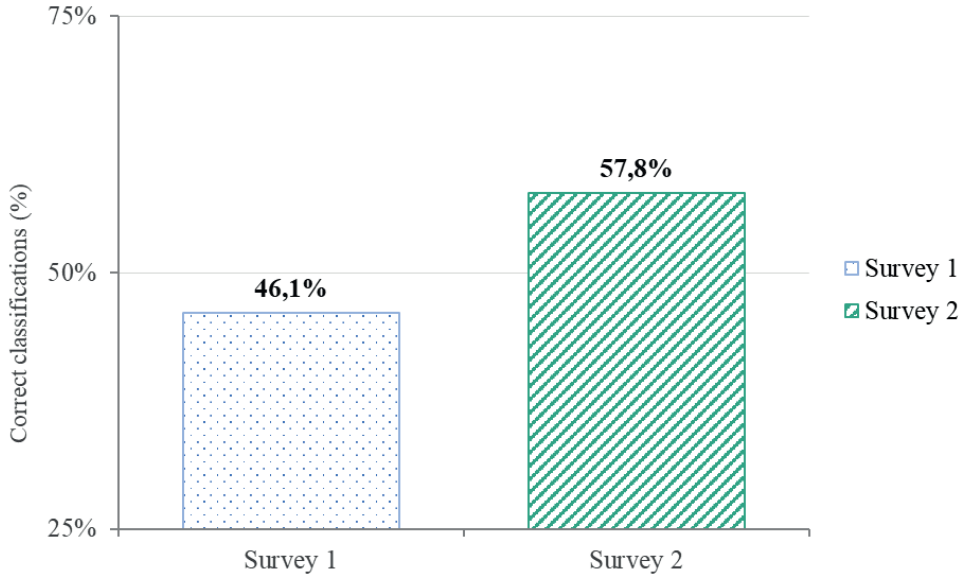

Figure 1. Classification Accuracy (n = 102)

S o u r c e: Own work.

Improvement from S1 to S2 (H3)

H3-A1: McNemar per-guess analysis (Figure 2a):

At the classification level, more trials improved from Survey 1 to Survey 2 (20) than declined (8). An exact one-tailed McNemar test was significant, p = .018. The odds of improvement were 2.50 times higher than the odds of decline (OR = 2.50), with a 95% CI [1.05, 6.56], indicating that improvement was more likely than decline.

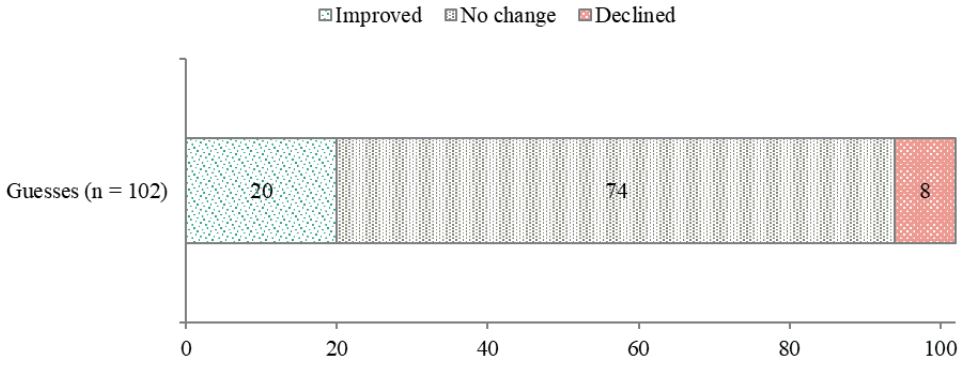

Figure 2a. Change in Classification: Guesses (n = 102)

S o u r c e: Own work.

H3-A2: McNemar per-student analysis (Figure 2b):

When analyzed at the student level using a "both-correct" indicator (1 = both items correct; 0 = otherwise), more students improved (10) than declined (3). An exact one-tailed McNemar test was significant, p = .046. The odds of improvement were 3.33 times higher than the odds of decline (OR = 3.33), with a 95% CI [0.86, 18.85]. Although the confidence interval includes 1, the directional test suggests improvement was more likely than decline. The number of students who classified both reviews correctly increased from 19 in Survey 1 to 26 in Survey 2.

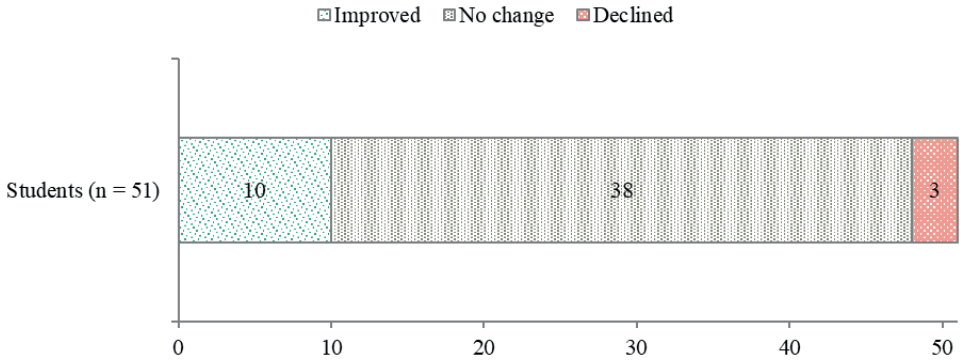

Figure 2b. Change in Classification: Students (N = 51)

S o u r c e: Own work.

H3-B: Wilcoxon per-student analysis (Figure 3):

A paired Wilcoxon signed-rank test indicated that students' Survey 2 accuracy scores (Mdn = 1.00, M = 0.58) were significantly higher than their Survey 1 scores (Mdn = 0.50, M = 0.46), W = 138.00, p = .037 (one-tailed). The Hodges–Lehmann pseudomedian shift (S2 − S1) was 0.25, with a two-sided 95% CI [0.00, 0.75]. The effect size was small-to-moderate (rank-biserial r = .45), suggesting students tended to achieve higher classification accuracy after engaging with the material.

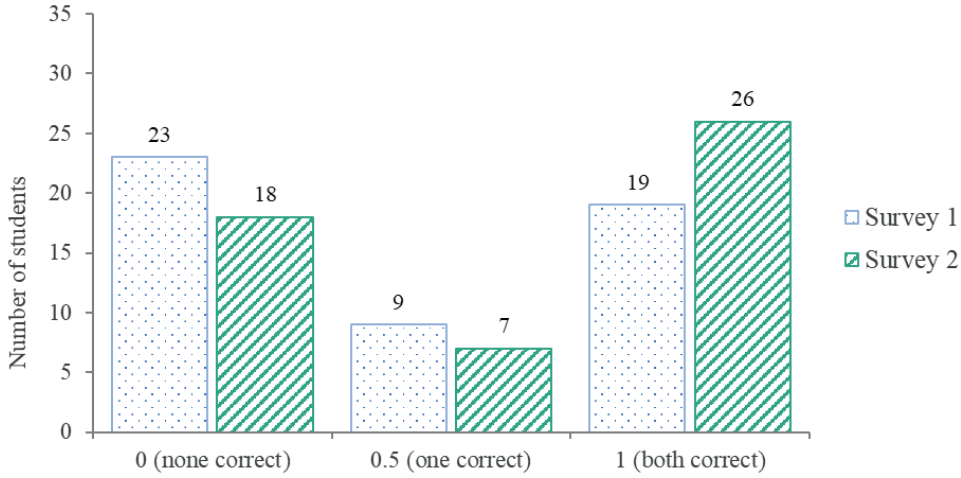

Figure 3. Distribution of Accuracy Score (N = 51)

S o u r c e: Own work.

Perceived Quality of Reviews

All respondents (REAL, N = 51):

No significant differences were found in comparing the reviews when determining whether they were human or LLM-generated on the basis of ease of reading, intuitive structure, or naturalness in either Survey 1 or Survey 2 (all BH-adjusted ps ≥ .221) (Table 1). Ease of response in Survey 2 was also non-significant (adj. p = .901).

Both-balanced subset (n = 38):

In the REAL analyses (Table 2), none of the dimensions reliably distinguished LLM from human reviews after correction for multiple comparisons (all adj. ps ≥ .389).
In the PERCEIVED analyses (Table 3), however, reviews classified as human were consistently rated more positively. In Survey 1, reviews labeled as human were judged easier to read (adj. p = .013), more intuitive (adj. p = .005), and more natural (adj. p < .001). In Survey 2, this same pattern held for ease of reading (adj. p = .028), intuitive structure (adj. p = .016), and naturalness (adj. p < .001), while ease of response did not differ significantly (adj. p = .326).

Both-correct in both surveys (n = 16; REAL = PERCEIVED):

For students who made no classification errors across both surveys (Table 4), human reviews were rated as more natural than LLM-generated reviews in Survey 1 (adj. p = .003). The trend of human reviews being rated higher on naturalness is also observed for Survey 2, but the difference is not considered statistically significant after correction (adj. p = .051). No statistically significant differences were observed for other metrics.

Table 1
Mean Ratings by True Source (All respondents, N = 51; REAL)

Survey	Dimension	AI M (SD)	Human M (SD)	Wilcoxon p	Adj. p
S1	Easy to read	4.33 (0.89)	4.18 (0.84)	.313	.395
S1	Intuitive structure	4.12 (0.84)	3.78 (0.90)	.074	.221
S1	Naturalness	3.69 (1.21)	3.43 (1.06)	.395	.395
S2	Easy to read	4.22 (0.88)	4.08 (0.87)	.457	.901
S2	Intuitive structure	3.84 (1.01)	3.82 (1.01)	.901	.901
S2	Naturalness	3.53 (1.25)	3.61 (1.15)	.777	.901
S2	Ease of response	3.90 (0.88)	3.84 (1.03)	.690	.901

Source: Own work.

Table 2
Mean Ratings (Both-balanced subset, n = 38) REAL

Survey	Dimension	AI M (SD)	Human M (SD)	Wilcoxon p	Adj. p
S1	Easy to read	4.34 (0.88)	4.16 (0.79)	.319	.479
S1	Intuitive structure	4.18 (0.90)	3.82 (0.95)	.130	.389
S1	Naturalness	3.71 (1.27)	3.47 (1.11)	.505	.505
S2	Easy to read	4.24 (0.88)	4.00 (0.87)	.295	.775
S2	Intuitive structure	3.84 (1.00)	3.74 (1.00)	.668	.775
S2	Naturalness	3.47 (1.29)	3.61 (1.15)	.691	.775
S2	Ease of response	3.89 (0.89)	3.84 (1.10)	.775	.775

S o u r c e: Own work.

Table 3
Mean Ratings (Both-balanced subset, n = 38) PERCEIVED

Survey	Dimension	Classified as AI M (SD)	Classified as Human M (SD)	Wilcoxon p	Adj. p
S1	Easy to read	4.03 (0.88)	4.47 (0.73)	.013	.013
S1	Intuitive structure	3.66 (0.97)	4.34 (0.78)	.003	.005
S1	Naturalness	2.74 (0.95)	4.45 (0.69)	< .001	< .001
S2	Easy to read	3.87 (0.84)	4.37 (0.85)	.021	.028
S2	Intuitive structure	3.42 (0.83)	4.16 (1.03)	.008	.016
S2	Naturalness	2.66 (0.91)	4.42 (0.86)	< .001	< .001
S2	Ease of response	3.76 (0.91)	3.97 (1.08)	.326	.326

S o u r c e: Own work.

Table 4
Mean Ratings (Both-correct in both surveys, n = 16; REAL = PERCEIVED)

Survey	Dimension	AI M (SD)	Human M (SD)	Wilcoxon p	Adj. p
S1	Easy to read	4.13 (1.02)	4.38 (0.72)	.372	.372
S1	Intuitive structure	3.75 (0.93)	4.06 (0.85)	.298	.372
S1	Naturalness	2.69 (1.08)	4.25 (0.86)	.001	.003
S2	Easy to read	4.00 (0.89)	4.00 (1.03)	.974	.974
S2	Intuitive structure	3.50 (0.89)	3.88 (1.09)	.417	.707
S2	Naturalness	2.69 (0.95)	4.25 (1.00)	.013	.051
S2	Ease of response	3.75 (0.93)	3.50 (1.32)	.530	.707

S o u r c e: Own work.

Personalization and Reception

The students' responses were overwhelmingly positive to the question concerning whether they found the personalization of the review process to their original text more interesting or engaging (Survey 1, 98% answered "Yes …"; Survey 2, 94% of students answered "Yes …").

Discussion

The findings of the study are substantive, addressing each of the research questions and inviting further investigation. There are three broad points to address here. The first is the qualitative assessment of the texts and how students perceived the material. The second is the corroboration of the consensus concerning detection at first interaction with the material, and the statistically significant improvement from the first to the second surveys. The third is the student reception of the exercise.

Qualitative Assessment

The findings suggest that the source of the reviews (REAL scenario; applies both to N = 51 and n = 38) did not reliably determine the participants' assessment of review quality (i.e., their assessments of ease of reading, intuitive structure, etc.). At the same time, the final classification decisions were strongly associated with the qualitative assessments (per the PERCEIVED scenario): in the both-balanced group (n = 38), reviews that students rated more positively across ease, intuitiveness, and naturalness were more likely to be labeled as human. However, among the stricter subgroup of students (n = 16) who achieved full classification accuracy, only assessments of naturalness differentiated human from LLM-generated reviews.

This might be taken to indicate that while most students relied on several perceived qualities when classifying, only naturalness consistently correlated with accurate classification of human versus LLM-generated reviews. Though the observation is limited to naturalness, the finding generally corroborates similar findings in studies involving human assessors (Howcroft et al., 2020; van der Lee et al., 2021), with humans associating what they perceive to be human writing with features like naturalness (Novikova et al., 2018; Clark et al., 2021). Perception of naturalness in this case is also likely affected by the content of the course, as the instructor will, in some sense, dictate what naturalness looks like in the specific context of academic writing. This also potentially pairs with the first point noted in the limitations section below.

Detection and Improvement

The study strongly corroborates the general consensus in the literature that human assessors cannot distinguish human- from LLM-generated material (Clark et al., 2021; Dugan et al., 2023; Wu et al., 2025). The participants were at random odds in classification for Survey 1. They were also at random odds in classification for Survey 2, though more accurate. This answers Q1 – students are not able to reliably detect LLM-generated content in academic writing contexts.

Despite the finding that students were at random odds in both surveys, there is a significant increase in accuracy between the two. That is, the improvement between the first and the second survey is statistically significant. This suggests an answer to Q2 – engaging with the material does lead students to assess it differently. The participants did not receive any training between these two surveys, which suggests that the improvement is due to the students needing to engage with the material of each review. While in the first survey, they were immediately asked for their impression of the text, the second survey was conducted as they submitted revised manuscripts and responses to each of the reviews. It would be appropriate to infer that having to read, reflect on, and prepare responses to each of the reviews gave the participants a better understanding of the material, and in the process, more accurately classified the texts as either human or LLM-generated.

The finding of randomness and of improvement between surveys connects substantively with both (1) – the literature on detection and (2) – the more general context of higher education at issue here. Each can be commented upon in turn.

(1) The initial finding that students are at random odds for both the first and second survey supports the above-mentioned consensus as potentially holding in this specific domain, joining other studies that examine specific material, like recipes, poetry, etc. (Köbis & Mossink, 2021; Soni & Wade, 2023). It also contrasts the singular study noted in the introduction that suggested that humans are at above-odds at detecting LLM-generated content in academic abstracts (Ma et al., 2023), though this may be due to the present study making use of newer models (ChatGPT-3 vs ChatGPT-4).

Observing that detection accuracy significantly increased in this diachronic setting without any training between surveys is also notable, particularly as participants did not receive training between surveys, in contrast to other diachronic studies conducted in this space (cf. Clark et al., 2021). The finding provides some impetus for further diachronic studies, with there being relatively few recent human-oriented studies on detection, with most recent work pursuing watermark, model-based, or at most human-aided techniques (Wu et al., 2025).

(2) Where higher education is concerned, we can draw a number of inferences here. The observation that students could not discern human from LLM-

generated writing carries with it the implication that LLMs may have reached a level of sophistication allowing them to be relied upon to produce material meeting the standards of academic writing courses. If students are not able to distinguish between material prepared by an academic writing instructor and material generated by an LLM, then presumably the LLM is able to produce material of comparable quality. Put differently, if students cannot discern the difference, then even if there were some subtle difference between the two sets discernible to professionals, it is evidently not something that students are in a position to observe. Barring some clear differentiating feature or error in the generated text, and barring an error in instruction leading to a susceptibility to misinterpret LLM-generated material for human-written material, for which there is no evidence here, there is no statistically significant difference between the human and LLM-generated reviews from the pedagogical perspective.

This pairs interestingly with the observation of improvement between surveys. The improvement points to a subtle counterindication to the above, as it implies students did ultimately discern something in the material, allowing them to better recognize that one of the texts was LLM-generated. No clear trend in the qualitative assessment of the material can be appealed to as an explanation, with naturalness only playing a partially significant role when paired with the actual source of the material, not proving statistically significant across every case, though correlated, and ultimately not necessarily a cause of the classification choice. The lack of helpful differentiation on this point in the qualitative component of the survey may suggest that students discerned something in the content itself of their reviews that led them to change their classification choices in the second survey. In light of that, one might cautiously conclude that while LLMs have reached a degree of sophistication to effectively mimic academic writing, they might not yet be able to seamlessly integrate into more complex, engaged exercises – for instance, providing reviews for student essays.

Student Reception

The surveys concluded with a question concerning the student's impression of the exercise, and specifically whether they found it more engaging to have personalized feedback. The data for the question provides a clear answer to Q3, such that students do like personalized feedback, and their impression is independent of their perception of whether the material was LLM-generated.

While this question could have led to a complex datapoint, with variation between underlying student assessments, ultimately, no further analysis is required. The overwhelming majority of students reported positively (98% for survey 1, 94% for survey 2). There was no significant difference in whether the students considered both reviews to be human, both to be LLM-generated, or a mix. This

provides some additional impetus for continuing to work on creative integrations of LLMs into course structures, insofar as students appear to welcome the personalization regardless of their perceptions of the underlying tools.

Limitations

The study is subject to a number of limitations that may be addressed in future research. The first is that the human reviews were prepared by a single instructor. This introduces some limitations for the generalization to human writing, as it depends upon a single source, albeit a native speaker and an experienced academic writing instructor in this case. The second is that the findings of the study are dependent upon the specific model used. As models are quickly developing, these findings will not necessarily be reflective of the state of affairs with future models. The third is population specificity, as the participants were university students at non-native English-speaking institutions. Given that these are language-based assessments, there might be variation in different classroom contexts with speakers with different linguistic backgrounds. The fourth is the absence of follow-up analysis exploring why participants classified the materials the way that they did, which would potentially shed additional light on shifts in assessments between the two surveys. Each of these limitations could be addressed in future research exploring similar research questions, new models, or different contexts.

Conclusions

The object of this study was to explore the potential for integrating LLM-generated material into the design of academic writing courses, approaching the issue through the lens of detection or evaluation. Three research questions were posed, addressing whether students could recognize LLM-generated content, whether sustained engagement with that material had any effect on recognition, and how students perceived the personalized material. The findings of the study suggest that students cannot reliably detect LLM-generated content, but that their accuracy significantly improved from the first survey to the second, after having reflected on and responded to the material at issue. It also showed that students responded positively to receiving tailored feedback, irrespective of the perceived source. These findings provide insight for the integration of current models into course designs, though potentially still seeing some limitations in more complex

exercises. Future research could build on this study design by introducing larger, more diverse samples, with further studies addressing the observed change in accuracy between surveys and further potential effects of reader engagement.

The code and data for this study may be found here: https://doi.org/10.5281/zenodo.17399250

The authors declare that they have no conflicts of interest.

References

Baidoo-Anu, D., & Owusu Ansah, L. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning (SSRN Scholarly Paper No. 4337484). *Social Science Research Network*. https://doi.org/10.2139/ssrn.4337484

Bedington, A., Halcomb, E. F., McKee, H. A., Sargent, T., & Smith, A. (2024). Writing with generative AI and human-machine teaming: Insights and recommendations from faculty and students. *Computers and Composition, 71*, 102833. https://doi.org/10.1016/j.compcom.2024.102833

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*(1), 43. https://doi.org/10.1186/s41239-023-00411-8

Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., & Smith, N. A. (2021). All that's "human" is not gold: Evaluating human evaluation of generated text. *arXiv*. https://doi.org/10.48550/arXiv.2107.00061

Clark, E., & Smith, N. A. (2021). Choose your own adventure: Paired suggestions in collaborative writing for evaluating story generation models. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3566–3575. https://doi.org/10.18653/v1/2021.naacl-main.279

Ding, L., & Zou, D. (2024). Automated writing evaluation systems: A systematic review of Grammarly, Pigai, and Criterion with a perspective on future directions in the age of generative artificial intelligence. *Education and Information Technologies, 29*(11), 14151–14203. https://doi.org/10.1007/s10639-023-12402-3

Du, H., Jia, Q., Gehringer, E., & Wang, X. (2024). Harnessing large language models to auto-evaluate the student project reports. *Computers and Education: Artificial Intelligence, 7*, 100268. https://doi.org/10.1016/j.caeai.2024.100268

Dugan, L., Ippolito, D., Kirubarajan, A., Shi, S., & Callison-Burch, C. (2023). Real or fake text? Investigating human ability to detect boundaries between human-written and machine-generated text. *Proceedings of the AAAI Conference on Artificial Intelligence, 37*(11), 12763–12771. https://doi.org/10.1609/aaai.v37i11.26501

Gabriel, I., Manzini, A., Keeling, G., Hendricks, L. A., Rieser, V., Iqbal, H., Tomašev, N., Ktena, I., Kenton, Z., Rodriguez, M., El-Sayed, S., Brown, S., Akbulut, C., Trask, A., Hughes, E., Bergman, A. S., Shelby, R., Marchal, N., Griffin, C., … Manyika, J. (2024). *The ethics of advanced AI assistants*. arXiv. https://doi.org/10.48550/ARXIV.2404.16244

Grassini, S. (2023). Shaping the future of education: Exploring the potential and consequences of AI and ChatGPT in educational settings. *Education Sciences, 13*(7), 692. https://doi.org/10.3390/educsci13070692

Howcroft, D. M., Belz, A., Clinciu, M.-A., Gkatzia, D., Hasan, S. A., Mahamood, S., Mille, S., Van Miltenburg, E., Santhanam, S., & Rieser, V. (2020). Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. *Proceedings of the 13th International Conference on Natural Language Generation*, 169–182. https://doi.org/10.18653/v1/2020.inlg-1.23

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

Köbis, N., & Mossink, L. D. (2021). Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. *Computers in Human Behavior, 114*, 106553. https://doi.org/10.1016/j.chb.2020.106553

Liu, W. (2024). A systematic review of automated writing evaluation feedback: Validity, effects and students' engagement. *Language Teaching Research Quarterly, 45*, 86–105. https://doi.org/10.32038/ltrq.2024.45.05

Ma, Y., Liu, J., Yi, F., Cheng, Q., Huang, Y., Lu, W., & Liu, X. (2023). *AI vs. human—Differentiation analysis of scientific content generation* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2301.10416

Novikova, J., Dušek, O., & Rieser, V. (2018). RankME: Reliable human ratings for natural language generation. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 72–78. https://doi.org/10.18653/v1/n18-2012

Ortiz-Bonnin, S., & Blahopoulou, J. (2025). Chat or cheat? Academic dishonesty, risk perceptions, and ChatGPT usage in higher education students. *Social Psychology of Education, 28*(1), 113. https://doi.org/10.1007/s11218-025-10080-2

Phokoye, S. P., Dlamini, S., Mthalane, P. P., Luthuli, M., & Moyane, S. P. (2025). A comprehensive review of ChatGPT in teaching and learning within higher education. *Informatics, 12*(3), 74. https://doi.org/10.3390/informatics12030074

Prunkl, C. E. A., Ashurst, C., Anderljung, M., Webb, H., Leike, J., & Dafoe, A. (2021). Institutionalizing ethics in AI through broader impact requirements. *Nature Machine Intelligence, 3*(2), 104–110. https://doi.org/10.1038/s42256-021-00298-y

Qian, Y. (2025). Pedagogical applications of generative AI in higher education: A systematic review of the field. *TechTrends*. https://doi.org/10.1007/s11528-025-01100-1

Rosas, R. A. P., & Rodríguez, E. J. A. (2024). Chatbots: The future of education? *International Journal of Engineering Pedagogy (iJEP), 14*(8), 107–119. https://doi.org/10.3991/ijep.v14i8.49715

Soni, M., & Wade, V. (2023). *Comparing abstractive summaries generated by ChatGPT to real summaries through blinded reviewers and text classification algorithms*. arXiv. https://doi.org/10.48550/arXiv.2303.17650

Stöhr, C., Ou, A. W., & Malmström, H. (2024). Perceptions and usage of AI chatbots among students in higher education across genders, academic levels and fields of study. *Computers and Education: Artificial Intelligence, 7*, 100259. https://doi.org/10.1016/j.caeai.2024.100259

Su, Y., & Wu, Y. (2024). *Robust detection of LLM-generated text: A comparative analysis*. arXiv. https://doi.org/10.48550/arXiv.2411.06248

Tzirides, A. O. (Olnancy), Zapata, G., Kastania, N. P., Saini, A. K., Castro, V., Ismael, S. A., You, Y., Santos, T. A. D., Searsmith, D., O'Brien, C., Cope, B., & Kalantzis, M. (2024). Combining human and artificial intelligence for enhanced AI literacy in higher education. *Computers and Education Open, 6*, 100184. https://doi.org/10.1016/j.caeo.2024.100184

van der Lee, C., Gatt, A., van Miltenburg, E., & Krahmer, E. (2021). Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language, 67*, 101151. https://doi.org/10.1016/j.csl.2020.101151

Wang, H., Li, J., & Li, Z. (2024). *AI-Generated text detection and classification based on BERT deep learning algorithm*. https://doi.org/10.48550/ARXIV.2405.16422

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity, 19*(1), 26. https://doi.org/10.1007/s40979-023-00146-z

Wu, J., Yang, S., Zhan, R., Yuan, Y., Chao, L. S., & Wong, D. F. (2025). A survey on LLM-generated text detection: Necessity, methods, and future directions. *Computational Linguistics, 51*(1), 275–338. https://doi.org/10.1162/coli_a_00549

Wu, J., Zhan, R., Wong, D. F., Yang, S., Yang, X., Yuan, Y., & Chao, L. S. (2024). DetectRL: Benchmarking LLM-generated text detection in real-world scenarios. *Advances in Neural Information Processing Systems, 37*, 100369–100401.

Yang, H., Kim, H., Lee, J. H., & Shin, D. (2022). Implementation of an AI chatbot as an English conversation partner in EFL speaking classes. *ReCALL, 34*(3), 327–343. https://doi.org/10.1017/S0958344022000039

Kamil Lemanek, Aleksander Leczkowski

Wdrażanie modeli językowych (LLM) w nauczanie pisania akademickiego: zaangażowane czytanie a materiały generowane przez modele językowe

S t r e s z c z e n i e

Włączanie dużych modeli językowych (LLM) do planów zajęć w szkolnictwie wyższym pozostaje otwartym i szybko rozwijającym się obszarem badań. Niniejsze badanie stanowi wkład w powstającą literaturę przedmiotu poprzez analizę tego, w jaki sposób studenci angażują się w interakcję z informacją zwrotną generowaną przez LLM oraz jak ją oceniają w ramach nauczania pisania akademickiego. Bazując na wcześniejszych pracach dotyczących integracji i wykrywania LLM, w badaniu zastosowano podejście diachroniczne w celu oceny, czy studenci potrafią wiarygodnie odróżnić recenzje sporządzone przez człowieka od tych wygenerowanych przez LLM oraz w jaki sposób zaangażowanie w materiał może wpływać na postrzegane klasyfikacje. Pięćdziesięciu jeden studentów studiów licencjackich otrzymało zarówno recenzje napisane przez człowieka, jak i wygenerowane przez LLM dla wypracowań, których byli autorami, wypełniając ankiety przed i po rewizji swoich prac. Wyniki wskazują, że studenci początkowo mieli losowe szanse na odróżnienie tekstu napisanego przez człowieka od tekstu wygenerowanego przez LLM, ale wykazali statystycznie istotną poprawę po zaangażowaniu się w materiał. Prawie wszyscy studenci zgłaszali pozytywne nastawienie do spersonalizowanych informacji zwrotnych, niezależnie od tego, czy postrzegali je jako generowane przez człowieka, czy przez LLM. Wyniki podkreślają zarówno potencjał pedagogiczny, jak i niejednoznaczność oceny integracji LLM w nauczaniu pisania.

S ł o w a k l u c z o w e: szkolnictwo wyższe; pisanie; detekcja; projektowanie kursów

Kamil Lemanek, Aleksander Leczkowski

Integración de los modelos de lenguaje grande (LLM) y la enseñanza de la escritura académica: lectura activa y material generado por LLM

Resumen

La integración de los grandes modelos de lenguaje (LLM) en el diseño de cursos en el contexto de la educación superior sigue siendo un campo de investigación abierto y en rápido crecimiento. Este estudio contribuye a la literatura emergente al examinar cómo los estudiantes interactúan con los comentarios generados por los LLM y cómo los evalúan en el marco de la enseñanza de la redacción académica. Partiendo de trabajos previos sobre la integración y la detección de los LLM, el estudio adopta un diseño diacrónico para evaluar si los estudiantes pueden distinguir de forma fiable entre los comentarios generados por personas y los generados por LLM, y cómo la interacción con el material puede afectar a la clasificación percibida. Cincuenta y un participantes de grado recibieron comentarios tanto generados por humanos como por LLM para los ensayos que escribieron, y completaron encuestas antes y después de revisar su trabajo. Los resultados indican que, inicialmente, los estudiantes tenían las mismas probabilidades de distinguir entre texto generado por humanos y por LLM, pero mostraron una mejora estadísticamente significativa tras la interacción. Casi todos los estudiantes manifestaron actitudes positivas hacia la retroalimentación personalizada, independientemente de si la percibían como generada por humanos o por un LLM. Los hallazgos subrayan tanto el potencial pedagógico como la ambigüedad evaluativa de la integración de los LLM en la enseñanza de la escritura.

Palabras clave: Educación superior; Redacción; Detección; Diseño de cursos

Камиль Леманек, Александр Лечковский

Интеграция больших языковых моделей (LLM) в преподавание академического письма: активное чтение и материалы, сгенерированные LLM

Аннотация

Интеграция крупных языковых моделей (LLM) в разработку учебных курсов в сфере высшего образования остается открытой и быстро развивающейся областью исследований. Данное исследование вносит вклад в формирующуюся научную литературу, анализируя, как студенты взаимодействуют с отзывами, сгенерированными LLM, и оценивают их в рамках обучения академическому письму. Опираясь на предыдущие работы по интеграции и распознаванию текстов, сгенерированных LLM, в исследовании используется диахронический подход для оценки того, могут ли студенты достоверно различать отзывы, написанные людьми и сгенерированные LLM, а также как взаимодействие с материалом может повлиять на их восприятие классификации. Пятьдесят один студент бакалавриата получил как отзывы, сгенерированные людьми, так и отзывы, сгенерированные LLM, на написанные ими эссе, заполнив анкеты до и после пересмотра своих работ. Результаты показывают, что изначально студенты случайно различали тексты, сгенерированные людьми и LLM, но продемонстрировали статистически значимое улучшение после взаимодействия с материалами. Почти все студенты сообщили о по-

ложительном отношении к персонализированной обратной связи, независимо от того, воспринимали ли они ее как созданную человеком или LLM. Полученные данные подчеркивают как педагогический потенциал, так и неоднозначность оценки интеграции LLM в обучение письму.

К л ю ч е в ы е с л о в а: высшее образование; письменная речь; выявление; разработка учебных курсов