Adv Physiol Educ 49: 486–495, 2025.
First published March 19, 2025; doi:10.1152/advan.00235.2024

american physiological society

ADVANCES IN PHYSIOLOGY EDUCATION

EDUCATION RESEARCH

Teaching in an Era of Generative Artificial Intelligence

Using aggregated AI detector outcomes to eliminate false positives in STEM-student writing

Jon-Philippe K. Hyatt,[1] Elisa Jayne Bienenstock,[2] Carla M. Firetto,[3] Elizabeth R. Woods,[1] and Robert C. Comus[1]

[1]*College of Integrative Sciences and Arts, Arizona State University, Tempe, Arizona, United States;* [2]*Watts College of Public Service and Community Solutions, Arizona State University, Tempe, Arizona, United States; and* [3]*Mary Lou Fulton College for Teaching and Learning Innovation, Arizona State University, Tempe, Arizona, United States*

Abstract

Generative artificial intelligence (AI) large language models have become sufficiently accessible and user-friendly to assist students with course work, studying tactics, and written communication. AI-generated writing is almost indistinguishable from human-derived work. Instructors must rely on intuition/experience and, recently, assistance from online AI detectors to help them distinguish between student- and AI-written material. Here, we tested the veracity of AI detectors for writing samples from a fact-heavy, lower-division undergraduate anatomy and physiology course. Student participants ($n = 190$) completed three parts: a hand-written essay answering a prompt on the structure/function of the plasma membrane; creating an AI-generated answer to the same prompt; and a survey seeking participants' views on the quality of each essay as well as general AI use. Randomly selected ($n = 50$) participant-written and AI-generated essays were blindly uploaded onto four AI detectors; a separate and unique group of randomly selected essays ($n = 48$) was provided to human raters ($n = 9$) for classification assessment. For the majority of essays, human raters and the best-performing AI detectors ($n = 3$) similarly identified their correct origin (84–95% and 93–98%, respectively) ($P > 0.05$). Approximately 1.3% and 5.0% of the essays were detected as false positives (human writing incorrectly labeled as AI) by AI detectors and human raters, respectively. Surveys generally indicated that students viewed the AI-generated work as better than their own ($P < 0.01$). Using AI detectors in aggregate reduced the likelihood of detecting a false positive to nearly 0%, and this strategy was validated against human rater-labeled false positives. Taken together, our findings show that AI detectors, when used together, become a powerful tool to inform instructors.

NEW & NOTEWORTHY We show how online artificial intelligence (AI) detectors can assist instructors in distinguishing between human- and AI-written work for written assignments. Although individual AI detectors may vary in their accuracy for correctly identifying the origin of written work, they are most effective when used in aggregate to inform instructors when human intuition gets it wrong. Using AI detectors for consensus detection reduces the false positive rate to nearly zero.

anatomy; physiology; undergraduate

INTRODUCTION

Learning anatomy and physiology (A&P) at the undergraduate level requires remembering facts and understanding physiological mechanisms. For students, facts accessed to build their understanding of A&P may come from a variety of avenues including, but not limited to, live lectures, primary (scientific research articles) and secondary (textbooks) sources, peer-to-peer communication (study groups), and the internet, providing additional/alternative lecture material through static websites and/or videos. Although the internet may compound information overload for students and threaten the limits of student attention span (1, 2), the internet provides a valuable avenue to confirm and clarify facts en route to their understanding of bodily function. Beyond the standard multiple-choice exam, writing assignments provide students an opportunity to demonstrate learning by practicing vocabulary and articulating A&P concepts in their own words. The emergence of excellent, easily accessible, user-friendly, and time-saving generative artificial intelligence (AI) tools to synthesize fact-heavy and difficult A&P concepts in a well-constructed and error-free written form has become an attractive lure for students (3).

AI can be a valuable tool to assist student learning (4–6) and communication, particularly if the student is a nonnative English speaker (7). The challenge for instructors is to

Correspondence: J.-P. K. Hyatt (jphyatt@asu.edu).
Submitted 2 December 2024 / Revised 13 January 2025 / Accepted 17 March 2025

teach students how AI can augment or streamline, but not replace, their efforts toward learning while maintaining fairness within the class. Students using AI to complete learning activities, assignments, or written essays, without actually completing them themselves, can lead to a lost opportunity for learning. Essentially, the time spent thinking and learning the content may be replaced by computations of an AI. Additionally, this can also lead to "illusions of understanding" (8), where students may not know as much as they think they do, subsequently leaving them underprepared for later material in the course, advanced courses, placement exams, and ultimately their intended careers. Instructors must consider the role of fairness in large-enrollment courses, both in how students are evaluated and graded as well as identification of those who have used AI against course policy. Detecting AI accurately is about more than an effort to identify instances of academic integrity violation efforts; it also serves as an opportunity to know if, when, and how much AI is being used so that instructors can make necessary adjustments to support students' mastery of the course content (e.g., revising the assignment, having conversations with students, ensuring students understand the learning costs associated with using AI).

If AI work is suspected on written assignments, then instructors may elect to verify student work with online AI detectors that generate a percent likelihood that the written work is "human" or "AI." Some AI detectors extend this evaluation further and provide a "mixture" category, denoting the likelihood that a portion of a submission involves a fusion of human- and AI-generated language. Similar to plagiarism checks when students turn in written work, some universities have chosen to integrate AI detectors into their learning management system (LMS) platforms to automatically check the degree of generative AI associated with the student essay. However, in any of these cases, there is no clear established cutoff for the percent likelihood that indicates AI use with any certainty. Consequently, one major concern for instructors (and administrators) is that a student's work is erroneously flagged as AI generated (e.g., a false positive), thereby falsely accusing the student of accessing help from AI tools.

Unlike plagiarism, where one can identify a word-for-word relationship between student work and a specific reference, AI-generated written material is arguably almost indistinguishable from actual writing (9, 10). Online AI detectors are generally believed to have a false positive rate of 2–10% (11, 12), which leaves a small probability that a false positive will be produced. Instructors and administrators may therefore conclude that AI detectors are generally untrustworthy or unhelpful, leaving concerned instructors with no resources to assist them in distinguishing student work from a well-crafted AI production. Here, we studied the veracity of AI detectors using short and verified human- and AI-generated STEM student writing samples from a large first-semester lower-division A&P course and compared their accuracy with that of human raters. We hypothesized that AI-generated writing would be identified with fewer errors by human raters than by AI detectors.

In parallel, we were interested in understanding whether AI was ubiquitously used by undergraduate STEM students. Earlier work reported that students use AI to find flaws in writing, analyze data and interpret results, improve multitasking, and help in writing in different languages (13). Given the inherent difficulty of learning A&P (14), which is compounded for some students whose first language is not English (15), we anticipated that A&P students would generally lack confidence in their ability to articulate A&P concepts shortly after learning the material. One goal here is to understand whether native and nonnative English-speaking students would rate the quality of AI work differently from their own.

METHODS

Participants

All work was reviewed and approved by the Institutional Review Board at Arizona State University (#STUDY00020548) and followed the experimental guidelines for human research participants according to the American Physiological Society. All participants provided written consent to participate in the research before beginning the first in-class exercise.

Undergraduate student participants were recruited from three sections of a lower-division human A&P course (BIO 201) taught by three instructors at a large public and Hispanic-serving university in the southwestern United States. The majority of the enrolled students were preprofessional health majors including, but not limited to, nursing, nutrition, kinesiology, medical studies, and biology. No prerequisite courses were required to enroll in this course. From our earlier work (16, 17), the general demographic representation of the students in this course is a 3:1 female-to-male ratio, ~19.5 ± 2.6 yr old, and 40% of students identifying as White. Approximately two-thirds of the students identify as non-Hispanic (64%), and one-third identify as Hispanic (36%). About one-third of students (32%) report being first-generation students.

All raw data were collected from September 9, 2024 to September 18, 2024. There were three requirements of the participants: writing an in-class essay to a prompt on plasma membrane physiology; using generative AI to write an essay to the same prompt; and the completion of a survey comparing the quality of each essay as well as their general use of AI. Students who completed all three requirements received extra credit.

For the in-class (human written) essay, students were provided a 20-min review on plasma membrane structure and function. Each review session was taught by the same instructor using identical material; all subject matter had been previously covered in each course section by different instructors. Bluebooks were provided to allow participants to take notes. In the same bluebook, participants were then instructed to hand-write an essay (150 ± 15 words) over the subsequent 20 min to the following prompt:

> Why do some substances diffuse across a cell's selectively permeable plasma membrane while others do not? Describe some ways in which those substances can either cross the membrane or get their message across the membrane and into the cell.

To ensure that essays were absent of outside/online influences, no personal electronic devices were allowed during the 20-min writing period. To mirror the way a student would compose a written essay outside of the classroom, all students were encouraged to use their hand-written bluebook notes to formulate their in-class essay. To further assist students with terminology and concepts, a review slide on membrane structure and function was projected in the lecture hall.

In the final 20 min of the in-class session, the participants were asked to copy their hand-written answer onto an electronic, non-cloud-based word processing medium (e.g., Word, Pages, Notes) and then upload that document onto the course's LMS. Students were asked to transcribe their written work into the word processing system verbatim, but they were permitted to correct minor spelling changes during transcription. Bluebooks were turned into the instructor as students exited the lecture hall; hand-written essays were securely stored and used to cross-check a participant's electronic version of the hand-written essay if necessary.

Two days after participants from each course's section completed and uploaded their hand-written essays, they were asked to identify and instruct a generative AI platform of their choosing to produce a 150-word essay to the same prompt used for the hand-written essay. The three AI platforms most used by the students were ChatGPT-4o (87.9%), Grammarly (3.2%), and Gemini (2.6%). Participants were asked not to alter the AI-written output and to copy it to a word processing document, which was then uploaded onto the course's LMS page. Participants were asked to save a copy of this AI-written essay.

From a total enrolled student starting population (Section 1: $n = 135$; Section 2: $n = 142$; Section 3: $n = 113$; total: $n = 388$), only the participants who uploaded both essays ($n = 260$) were then sent a link to a survey asking them to assess and compare the quality of both their and the AI-generated responses (Fig. 1); the survey concluded with questions asking the participants about their general AI use. Of those who completed the survey ($n = 190$), only essays submitted by the participants who gave consent for the use of their essays in internet AI detectors were included ($n = 174$; Fig. 1). Of the survey respondents, approximately 75% and 18% spoke English and Spanish as their first language, respectively. The remainder of the first languages spoken by survey participants included Arabic, Korean, Mandarin, Japanese, Dari, Urdu, Tagalog, and Turkish, with each constituting ~1–2% of the participant pool.

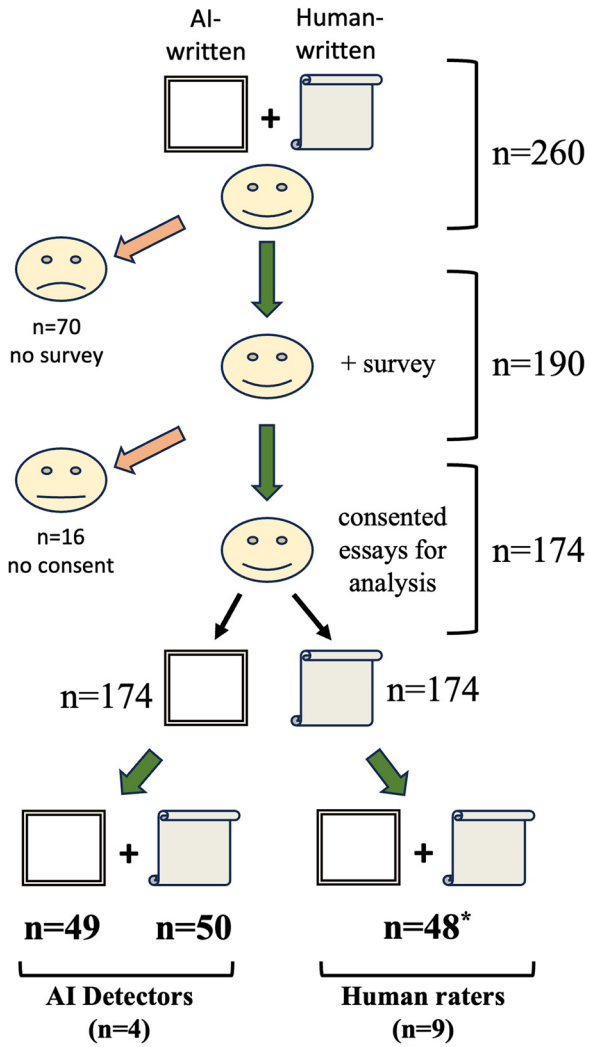

Figure 1. Schematic showing participant/essay sampling and attrition examining the veracity of artificial intelligence (AI) detectors and the accuracy of human raters in identifying human- vs. AI-written work at a large public university. *The essays randomly chosen for human raters were different from the essays randomly chosen for AI detector analysis.

Essay Classification by Human Raters

Three groups of human raters were asked to classify randomly chosen essays as either human or AI generated. Each group consisted of a faculty member, a graduate teaching assistant (TA), and an undergraduate TA. All human raters worked independently. Each rater was provided 20 essays that did not overlap with the essays analyzed by AI detectors. Human raters were asked to read all essays before assigning a score (0 = human; 1 = AI). To assess interrater reliability, 6 essays (3 human and 3 AI generated) of each group's 20 essays were identical. The remaining 14 essays added to each group were randomly chosen from the essay pool to eliminate any potential pattern recognition by the human raters. Once the essays were classified by human raters, they were then categorized as true positive, true negative, false positive, or false negative (Fig. 2).

Essay Classification by AI Detectors

Human- and AI-generated essays from 50 randomly chosen participants were used to determine the veracity of online AI detectors (100 essays total). One AI-generated essay file was corrupted and unreadable; it was subsequently removed. Four AI detectors were used to assess the origin of each essay, including GPTzero, Originality AI, Detect GPT, and Copyleaks. These AI detectors were chosen based on earlier reports (11, 12, 18), availability, and ease of navigation for the investigators. Essays were uploaded blind onto these AI detector platforms. Generally, AI detectors output a score (out of 100) for the likelihood that a writing sample is AI, human, or a mixture of both. Outputs were transferred to a database and compared to the known origin of each essay.

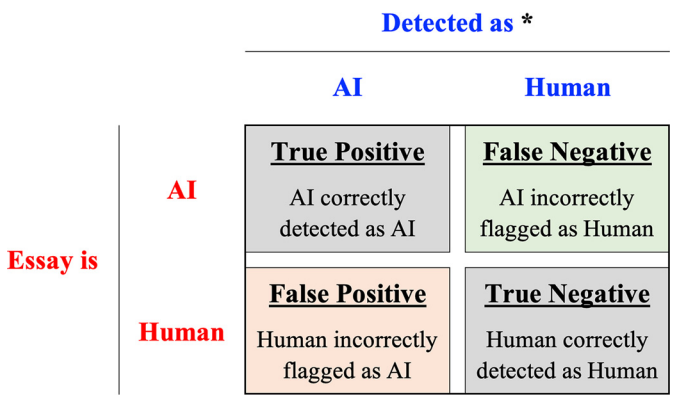

Figure 2. Classifications used for essays following scoring by artificial intelligence (AI) detector or human raters. *Essays that were assessed by AI detectors defaulted as human written if they were scored ≤80% likelihood of AI origin.

Essay classification by the AI detectors was the same that was used for human raters (Fig. 2); a threshold score of 80% AI was employed. For example, if an essay was detected as ≥80% AI and it was actually AI, then it was coded a true positive. However, if an essay was detected at ≥80% AI and was actually human, then it was classified as a false positive. Any essay that did not meet that 80% AI threshold was coded as "human." If the human-coded essay was AI generated, then it was coded as a false negative; if it was actually human generated, then it was coded as a true negative. While we know of no other established cutoff value that is commonly used by others, the ≥80% was chosen because we previously used that cutoff in a recent large online course that gave a reasonable "benefit of the doubt" that an essay was student written while capturing the majority of the essays that were "cut-and-pastes" from an AI program.

Survey

The purpose of the survey was to measure a student's impressions of the AI-generated output in comparison to their own writing along multiple dimensions. First, students were asked to read and review their own essay and evaluate it on a variety of characteristics: content, style, grammar/punctuation, and answering the prompt, based on a five-point scale (i.e., Poor to Excellent). Then, they evaluated their AI-generated essay in the same manner. To determine whether aggregated assessment of each essay was consistent at the individual level, the assessments of AI-generated essays were subtracted from the assessments of self-generated essays for each of these categories and statistically compared. Finally, the participants were asked to directly compare the two essays, identifying whether their essay was better, the AI essay was better, or they were equal. A final set of items focused on the extent to which students reported prior experience using AI.

Statistics

The Fisher exact test was used to test the significance of AI detector and human rater accuracy. AI detector and human rater classification rates were compared by a Student's two-tailed *t* test assuming unequal variance. Interrater reliability and survey data were analyzed with R: A Language and Environment for Statistical Computing (19) and RStudio (20). Interrater reliability was tested using the Fliess kappa statistic (21) for greater than two raters. Frequency and contingency tables were computed with Base R and the packages "epiDisplay" and "gmodels" (22, 23). Chi-square

***Essay 130**: Every cell has a selectively permeable membrane that is used to regulate which substances are able to enter and exit the membrane. The permeability of a substance depends on size, charge/polarity, and concentration gradient. Small molecules have a much higher permeability and are able to pass through the phospholipid bilayer easier than larger molecules. Nonpolar and neutrally charged molecules are able to pass through easier than polar and charged molecules, and areas where there is a favorable concentration gradient help to allow molecules in as well. For bigger, polar, and charged molecules, it may take facilitated diffusion, or active transport to get their message across the membrane and into the cell. Facilitated diffusion uses a carrier to bring the molecule in and active transport uses energy or ATP. If a molecule is still not able to pass through, a protein receptor on the membrane can send a signal and relay the message.*

***Essay 213**: The plasma membrane is composed primarily of a phospholipid bilayer interspersed with proteins. The bilayer's hydrophobic interior creates a barrier to most water-soluble and charged substances, while its outer and inner hydrophilic surfaces interface with the aqueous environments inside and outside the cell. This dual nature of the membrane contributes to its selective permeability. Substances that cannot cross the plasma membrane directly still have ways to influence cellular activities through various mechanisms. The selective permeability of the plasma membrane ensures that essential substances can enter and exit the cell while maintaining internal stability. Small, nonpolar molecules can diffuse directly through the lipid bilayer, whereas larger, polar molecules and ions require specialized transport mechanisms. Substances that cannot cross the membrane directly can still influence cellular activities through active transport, vesicular transport, and receptor-mediated signal transduction. These mechanisms collectively enable cells to maintain homeostasis and respond to their environment effectively, highlighting the intricate and dynamic nature of cellular function.*

***Essay 267**: A substance's ability to cross through a cell's selectively permeable plasma membrane depends on its polarity, size or concentration gradient. Molecules such as gasses that are super small and nonpolar, are highly permeable, meaning that they can pass through with little to no effort. This is also an example of a passive transport called simple diffusion. On the other hand, molecules such as ions, atoms with a positive or negative charge, are impermeable to the membrane and cannot pass through unless assisted. These ions need help from proteins such as the sodium potassium pump, as well as energy from ATP in order to successfully pass through the membrane. For these atoms, the size does not matter, as the charge is what determines its ability to pass. Overall, nonpolar molecules will have the easiest time passing through, while atoms with a charge take more energy.*

Figure 3. Representative examples of human (*essays 130* and *267*)- and artificial intelligence (AI) (*essay 213*)-written answers in response to the prompt on plasma membrane anatomy and physiology (see METHODS). All essays shown were assessed by human raters and were classified as follows: *essay 130*: true negative; *essay 213*: true positive; *essay 267*: false positive. *Essay 267* was then classified by AI detectors for additional analysis (see Fig. 7).

statistics were computed with the package "freqtables" (24) for univariate frequencies of student responses to questions regarding their assessments of the quality of essays. Base R and the CrossTables() command from "gmodels" (23) was used to generate contingency tables and for generating the chi-square statistic for testing difference in the use of AI for native English speakers versus others. Values are reported as mean ± standard error (SE). Significance levels were set at $P < 0.05$; however, differences of $P < 0.01$ and trends ($P = 0.06–0.1$) are noted when detected.

RESULTS

Accuracy of Human Raters

The essays that were written by AI and humans were similar, containing general fact-based information that directly answered the prompt (Fig. 3). True positive and true negative rates for human raters were 84.6 ± 6.3% and 95.0 ± 2.1%, respectively, whereas false positive and false negative rates for human raters were 5.0 ± 2.1% and 15.4 ± 6.3%, respectively (Fig. 4). There were no differences in false positive rates for faculty ($n = 3$; 3.0 ± 3.0%), graduate TAs ($n = 3$; 8.2 ± 4.5%), and undergraduate TAs ($n = 3$; 3.7 ± 3.7%), respectively ($P > 0.05$; Fig. 5). There was no difference in the false negative rates between faculty (14.8 ± 9.8%) and undergraduate TAs (31.3 ± 11.6%; $P > 0.05$), although graduate TAs' false negative rate (0%) was significantly lower than that of undergraduate TAs ($P < 0.05$). Interrater reliability between human raters was determined to be in "substantial agreement" ($k = 0.62$; Ref. 25).

Veracity of AI Detectors

The best-performing three of four AI detectors (Copyleaks, GPTzero, Originality AI) were largely accurate in their classification of STEM essays and did not statistically significantly differ from the accuracy of the classification rates of human raters ($P > 0.05$). Collectively, the best-performing AI detectors had true positive and negative rates of 93.9 ± 2.4% and 98.7 ± 0.7%, respectively, and false positive and negative rates of 1.3 ± 0.7% and 6.1 ± 2.4%, respectively (Fig. 4). There were fewer false positives and negatives classified by Copyleaks (0 and 5 detected, respectively), GPTzero (1 and 2 detected, respectively), and Originality AI (1 and 1 detected, respectively) than by Detect GPT (9 and 21 detected, respectively). However, when aggregating across all four AI detectors, including Detect GPT, the false positive and negative errors increased to 5.5 ± 4.2% and 15.3 ± 9.3% (Fig. 5). Of note, the false negative rate was higher than the false positive rate for both human raters and AI detectors.

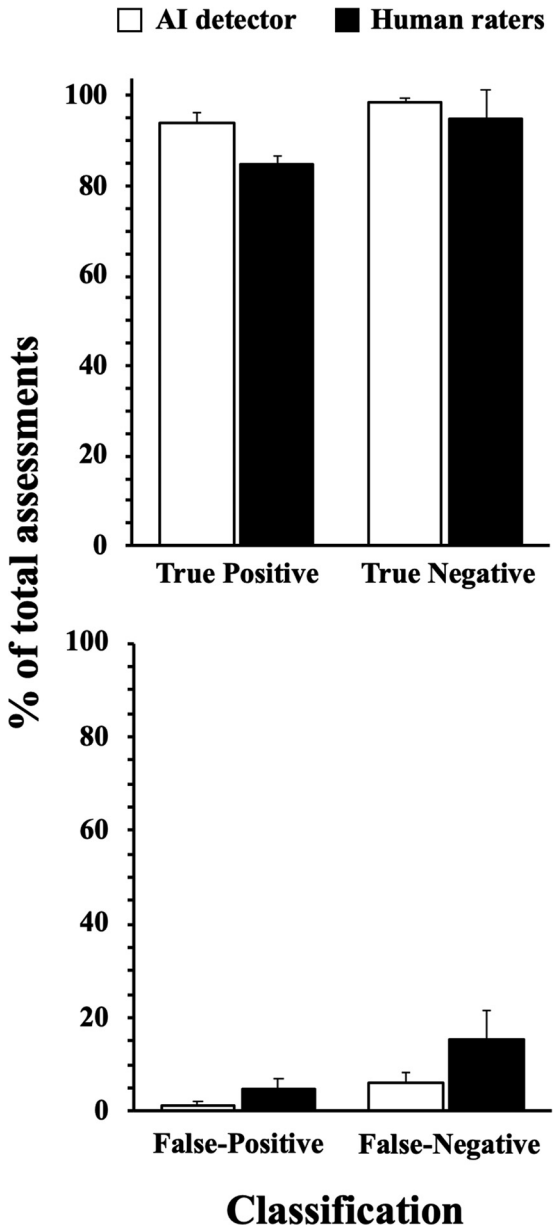

Figure 4. Detection rates for human- and artificial intelligence (AI)-written essays as determined by the best-performing online AI detectors ($n = 3$) and human raters ($n = 9$). *Top*: classification rates for essays that were correctly identified as AI (true positive) or human (true negative) written. *Bottom*: classification rates for essays that were incorrectly labeled as AI (false positive) or human (false negative) written. One AI detector was inconsistent in its classifications (e.g., high false positive/negative rates) compared to the other AI detectors used in this study and was removed from the calculations shown here. Values are means ± SE.

Aggregated AI Detector False Positive Rates

Using the individual AI detector false positive detection rates, we calculated the aggregated rates if the detectors were employed in tandem or as a triad (Fig. 6). Of the nine false positives identified by Detect GPT, there was only one overlapping false positive essay with GPTzero and another with Originality AI, reducing the likelihood of a false positive detection from a Detect GPT-GPTzero or a Detect GPT-Originality AI pairing to 0.36%. No false positives were identified by Copyleaks; using this AI detector in tandem with any of the other three detectors would lower the joint probability of finding a false positive to (near) 0%. Using AI detectors in pairs (Fig. 6) is impractical because there would be an inherent disagreement if one AI checker classifies a false positive and the other does not. However, if the AI detectors are used as a triad (Fig. 6), then the joint probability of detecting a false positive is 0% when a triad includes Copyleaks. The joint probability of detecting a false positive when using a GPTzero-Detect GPT-Originality AI triad is 0.0073%.

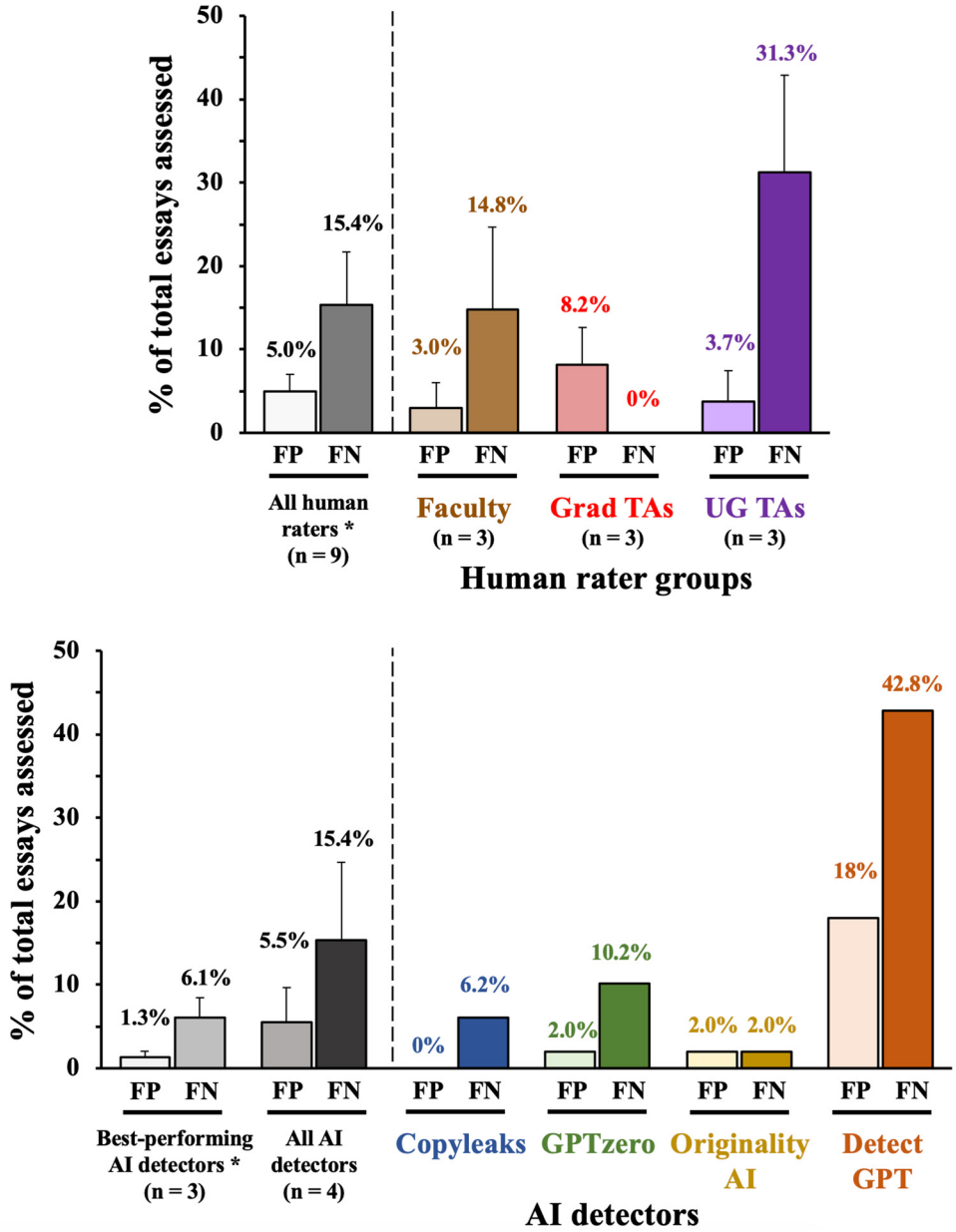

Figure 5. False positive (FP) and false negative (FN) classification rates (means ± SE) for human raters (*top*) and artificial intelligence (AI) detectors (*bottom*). Faculty and graduate or undergraduate teaching assistant (TA) FP/FN classification rates are shown as an aggregate (mean) or averaged. AI detector FP/FN classification rates are shown as a mean or individually. Mean values were calculated from 3 AI detectors (Copyleaks, GPTzero, and Originality AI) or with all 4 detectors. *FP and FN data are the same as shown in Fig. 4.

Using Aggregated AI Detector Outcomes to Reassess Instructors' False Positive Essays

To test the idea that AI detectors can help to inform instructors, all four human-written essays that were originally classified as false positives by human raters were uploaded and scored by the AI detectors used in this study (Fig. 7). In contrast to the original AI classifications by human raters, the four false positives were determined to be human-written essays when factoring in the aggregate AI recommendation (Fig. 7).

Participant Views on Self- and AI-Written Answers

Survey analysis revealed that, on average, the participants thought the AI essay was a stronger product than their own ($P < 0.01$; Table 1). For example, the percentage of total survey respondents rating their essays as Average or Good was 86.8%, 80%, 73.2%, and 75.8% for content, style, grammar/punctuation, and how well their essay answered the prompt, respectively. The percentage of total survey respondents rating the AI essay Good or Excellent for content, style, grammar/punctuation, and how well the AI essay answered the prompt was 91.6%, 84.2%, 92.1%, and 89.4%, respectively. These aggregated opinions were confirmed at an individual level for all categories: the most frequent values for the difference between ratings of self- and AI-generated essays were 0 and −1 (e.g., AI was rated 1 category better) for all comparisons ($P < 0.01$; data not shown). In addition, when comparing the essays directly, a majority of total survey respondents thought that the AI essay was the better answer (63.2%), was better written (75.1%), had greater depth (68.1%), and had fewer grammatical errors (57.4%) ($P < 0.05$; Table 1). Preliminary qualitative analysis of human- and AI-written content, however, showed no differences in the use

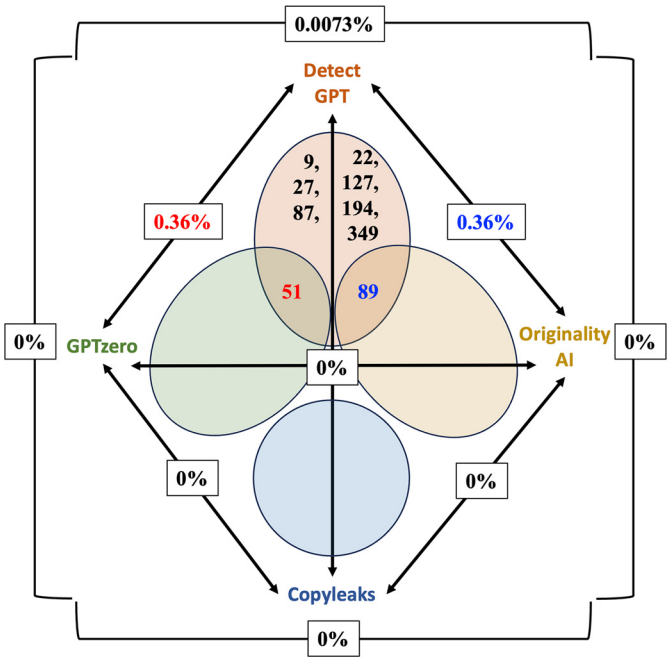

Figure 6. Probability of detecting false positives when using artificial intelligence (AI) detectors in tandem (arrows) or as a triad (brackets). A Venn diagram shows which AI detectors classified the same essay(s) as a false positive. Values within ellipses indicate essay numbers. One AI detector (Copyleaks) identified no false positives and is therefore shown detached. Importantly, the 0% value is a calculation based on the observed data and not an absolute/mathematical zero.

of A&P-related terminology within their write-ups (data not shown). There was a slightly higher tendency for native English speakers to judge the content of their essay as equal in quality to the AI-generated product; conversely, nonnative English speakers tended to rate the AI product as better ($P = 0.07$; data not shown).

General Use of AI by Undergraduate Students

A majority of total survey respondents indicated that they did not use AI to assist them with writing essays/papers (78.4%), writing essay/paper outlines (56.4%), writing emails (82.1%), creativity (50.6%), and assistance with mathematics (58.4%) (Table 2). However, a majority of total survey respondents admitted using AI to assist them with generating ideas to jump start them on assignments (63.7%) and finding errors in their work such as grammar or punctuation (56.3%) (Table 2). Generally, we found no differences in general AI assistance between native and nonnative English-speaking survey participants, with the exception that nonnative English-speaking survey participants used AI to assist in crafting emails more often than native English-speaking individuals ($P < 0.05$; Table 3).

DISCUSSION

The goal of this study was to determine accuracy of AI detectors and human raters in distinguishing between human- and AI-written work centered on a fact-heavy A&P-specific topic. There are two principal findings of this study. First, when AI detectors are used together, as a tandem or a triad, the likelihood of producing a false positive outcome (e.g., labeling a student-written paper as AI) is practically zero when the student writing is limited in scope and length and the AI writing is unaltered upon submission. Second, using AI detectors in aggregate will reduce, if not eliminate, human/instructor-labeled false positives. In agreement with earlier work, we also discovered variability in the accuracy of different AI detectors (11, 18); however, contrary to our hypothesis and others (10), we did not find a statistically significant difference in detection accuracy between high-performing AI detectors and human raters (Fig. 4). Taken together, our findings suggest that aggregating AI detectors could potentially detect false positives in STEM student writing by assisting instructor grading and intuition.

As demonstrated here, AI detectors, when used in aggregate, can provide additional information to the instructor to draw a conclusion (Fig. 7). Interestingly, Copyleaks, which did not detect any false positives from the original batch of human essays (Fig. 5 and Fig. 6), incorrectly labeled *essay 267* as AI, confirming that no single AI detector is perfect (Fig. 7). This observation underscores that, although there may be an absence in detection/perception of false positives by AI detectors (Fig. 5 and Fig. 6), this absence is not equal to a mathematical zero (26). In short, there is no absolute guarantee that any single AI detector will never detect a false positive. Furthermore, it is advised that an odd number of AI detectors should be used by instructors in the event that the AI detector consensus/tally score is tied.

There have been a number of experimental strategies used to test the abilities of AI detectors and human raters to distinguish AI and human writing. Expectedly, vendors for AI detectors self-advertise high accuracy rates; in parallel, there are also many anecdotal trial-and-error accounts of individuals testing detection rates with a cohort of known human- or AI-written samples (12). More formal studies testing the accuracy of AI detectors, however, have relied on a wide range of AI- and human-generated writing samples. AI writing has been used from entire published articles (18), limited passages or abstracts from publications (11, 27), or answers to an original prompt (9, 12). Human-generated samples have been convenience-sampled from student submissions covering a variety of topics from past coursework (9, 28), passages from self-written work (11), and published scientific opinion pieces (9). Interestingly, Perkins et al. (10), for example, submitted assignments fabricated by AI and asked instructors to identify any AI-developed work: AI flagged 91% of AI-generated work, whereas instructors only found AI submissions ~55% of the time.

It should be noted that an on-ground A&P course was used for this study because it gave the investigators control of the in-person writing environment that could not be achieved with, for example, an online-only A&P course. However, the findings presented here could be applied to both instructional types. Our experimental approach also employed a writing exercise with limited scope, depth, and length for both the AI and human writing samples, possibly somewhat restricting the wording, phrasing, and content variability in all essays. Our methodological approach attempted to mirror a realistic writing assignment used in lower-division A&P classes to test comprehension of difficult A&P concepts but an assignment that would also allow students to practice and

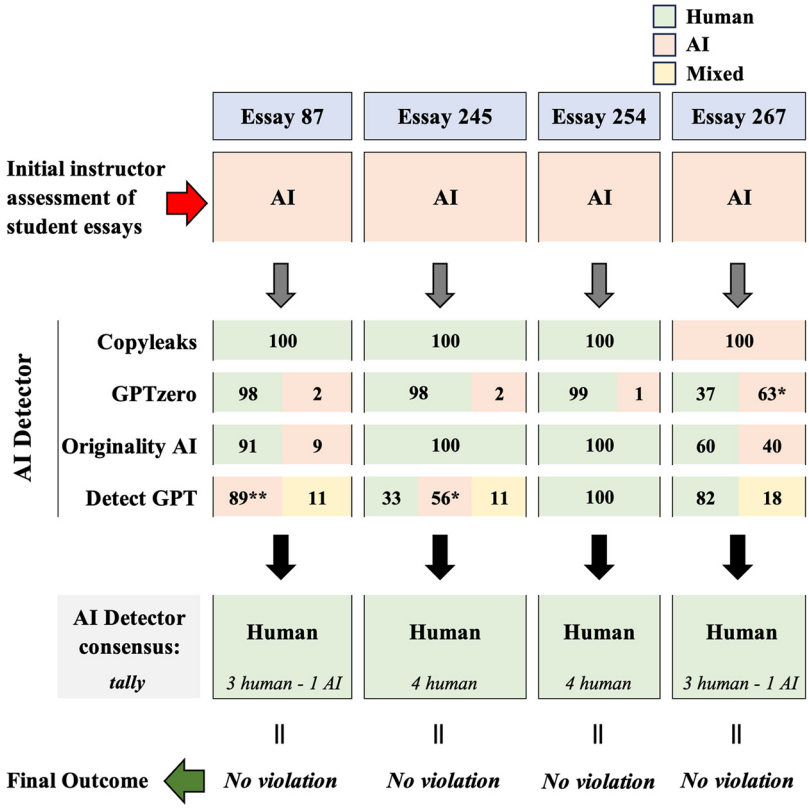

Figure 7. Using artificial intelligence (AI) detectors in aggregate to assist instructors in determining the origin of STEM student-written material. Four essays that were categorized as false positives by human raters were used as case studies. AI detectors assess writing out of a score of 100% and assess the likelihood that the writing sample is of human (green), AI (red), or mixed (yellow) origin. A positive AI detection was considered ≥80%; otherwise the essay defaulted as human written (e.g., giving the student the benefit of the doubt). After each essay was scored by each AI detector, a tally was made determining the AI detector consensus to determine a final outcome of the essay. *Defaulted as human because the AI score did not achieve the ≥80% threshold; **essay rated as AI because the score is above the 80% threshold.

contextualize STEM language into their own words. This experimental setup, however, may have allowed human raters to detect AI-generated material with a higher degree of accuracy than in earlier reports (10, 28). Interestingly, the human raters who were used in this study were from different disciplines and backgrounds with variability in time served as instructors. The faculty were from A&P, social science, and education, with the latter two having experience in biology and biology education research. The graduate TAs were currently, or previously, enrolled in an education-related Ph.D. program, and all had some formal biology-based education (i.e., high school concentration, undergraduate major/minor, master's degrees). All undergraduate TAs were majoring in science-specific fields. Although it is unclear whether the variability in backgrounds played a role in both the number of false positive and false negatives from each group as well as the strategies employed to identify the AI essays, this may warrant further investigation in follow-up studies using a larger sample of human raters.

Table 1. *Percentage of total responses by category for questions on perceptions of essay quality*

	Poor	Below Average	Average	Good	Excellent
After rereading YOUR* essay, how would you rate YOUR RESPONSE to the essay prompt in terms of:					
Content (information/facts)	1.6	4.7	36.3	50.5	6.8
Style (how well written)	2.1	9.5	46.8	33.2	8.4
Grammar/punctuation	1.6	8.9	30.0	43.2	16.3
Answering the prompt	0	4.2	24.2	51.6	20.0
After rereading the AI essay, how would you rate the AI RESPONSE to the essay prompt in terms of:					
Content (information/facts)	0	0	8.4	41.1	50.5
Style (how well written)	0.5	1.1	14.3	46.6	37.6
Grammar/punctuation	0.5	0.5	6.8	29.5	62.6
Answering the prompt	0	0	10.5	36.8	52.6
	YOUR Essay	**Equal**	**AI Essay**		
In comparing YOUR response vs. the AI response, which:					
is the better answer?	12.1	24.7	63.2		
is better written?	12.7	12.2	75.1		
goes into greater depth?	20.2	11.7	68.1		
has fewer factual (content) errors?	13.2	43.2	43.7		
has fewer grammatical (style) errors?	10.5	32.1	57.4		

Values shown are percentage of total respondents ($n = 190$). AI, artificial intelligence. *Participant-written essay.

Surveys

Although only a small number of students had indicated using AI assistance before this study (Table 3), students viewed AI-written essays as better/stronger than their own work (Table 1). The survey results imply that students' insecurity in their writing abilities to articulate difficult A&P concepts may be one compelling factor for AI use. Although generative AI has been shown to be useful for individuals learning English as a second language (7), Liang et al. (29) note that AI detectors exhibit bias against nonnative English speakers. We did not find that nonnative English-speaking undergraduate students reported prior use of AI differently from native English speakers, except for assistance in crafting emails (Table 3). Future research should examine the extent to which there are differences in quality and terminology between the human- and AI-written essays and whether differences in accuracy rates may be present for native versus nonnative English speakers.

Limitations

This study used writing samples that were written entirely by either humans or AI. Study participants wrote on a narrow topic and in a limited time frame. Although it is not uncommon to assign a writing exercise on a specific A&P concept, this study did not comprehensively assess how well the AI detectors or human raters could identify hybrid essays that fused AI- and human-generated language. It does appear that the four AI detectors used in this study had difficulty with human-AI hybrid essays. To test this, we exchanged sentences between the AI- and human-written essays from two participants. From this small exercise, we can report that AI detector accuracy declined as human- and AI-generated sentences became more intertwined (data not shown), and further experimentation is required to test, quantify, and validate these observations. Furthermore, AI generation and AI detectors are constantly improving; it is difficult to know the generalizability of these findings into the future as AI generation and AI detection evolve. However, assuming that AI generation and AI detection both improve at the same rate, our findings, and, more importantly, our recommendation that individuals and academic institutions should use multiple AI detectors to reduce false positive rates, is not unreasonable for the foreseeable future. Finally, only four AI detectors were used in this study; the use of AI detectors in aggregate to inform instructors should be conducted with an odd number of detectors in the event a tally score is tied (Fig. 7).

Table 2. *General AI use by lower-division anatomy and physiology undergraduate participants*

	Yes	No	Significance Level, *P* Value
Other than this study, have you used AI to assist you for your college work/assignments in:			
Writing essays/papers	21.6	78.4	<0.01
Writing essay/paper outlines (but not writing the paper itself)	43.7	56.4	0.08
Writing emails	17.9	82.1	<0.01
Helping you with ideas to get you started on assignments	63.7	36.3	<0.01
Helping you with creativity	49.5	50.6	0.93
Helping you with math	42.6	58.4	<0.05
Finding errors in your work (e.g., grammar, punctuation)	56.3	43.7	0.08

Values are percentage of total survey respondents. AI, artificial intelligence.

Table 3. *General AI use by first language*

	Yes (Indicates Prior AI Use)			No (Indicates No prior AI Use)
	Nonnative English Speakers	Native English Speakers	Total	Total
Essays	**11**	**30**	**41**	**149**
	23	*21*	*22*	*78*
Outlines	**22**	**61**	**83**	**107**
	46	*43*	*44*	*56*
Email*	**14**	**20**	**34**	**156**
	29	*14*	*18*	*82*
Math	**17**	**64**	**81**	**109**
	35	*45*	*43*	*57*
Ideas	**30**	**91**	**121**	**69**
	65	*64*	*64*	*36*
Creativity	**24**	**70**	**94**	**96**
	50	*49*	*49*	*51*
Total	**48**	**142**	**190**	**190**
	25	75	100	100

Boldface numbers are absolute values. Italicized numbers are percentages of total survey participants. AI, artificial intelligence. *Significantly different [$P < 0.029$; confidence interval (CI) = 0.171–0.955] between nonnative and native English speakers.

Summary

AI has clear benefits and drawbacks within higher education. As AI becomes more commonplace within academia, instructors require tools to help them accurately distinguish student- and AI-generated written material and to help instructors avoid falsely classifying student work as AI written while also accurately identifying instances of AI use so that they can support students in ways that will lead to better learning outcomes. From our present findings, we recommend that instructors, or administrators, use at least three AI detectors to eliminate the likelihood of false positives. The likelihood that any two AI detectors that we used identified the same false positive was <0.4%, but when three were used the false positive detection rate was reduced to 0–0.0073%, even though this detection rate will never reach a mathematical/theoretical zero. To assist instructors and to maintain fairness within classrooms, we recommend that institutions embed three or more AI detectors within their LMS, providing instructors with an important and necessary resource for their courses.

DATA AVAILABILITY

Data will be made available upon reasonable request.

ACKNOWLEDGMENTS

The authors thank Samantha Santos, Lin Yan, Emmanuel Adeloju, Caitlyn Green, Jacob Edwards-Scothern, and Sankit Patel for essay evaluations. In addition, the authors thank Monica Gaughan for survey development assistance.

DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the authors.

AUTHOR CONTRIBUTIONS

J.-P.K.H., E.J.B., and C.M.F. conceived and designed research; J.-P.K.H., E.J.B., E.R.W., and R.C.C. performed experiments; J.-P.K.H., E.J.B., E.R.W., and R.C.C. analyzed data; J.-P.K.H., E.J.B., and C.M.F. interpreted results of experiments; J.-P.K.H. prepared figures; J.-P.K.H. and C.M.F. drafted manuscript; J.-P.K.H., E.J.B., C.M.F., E.R.W., and R.C.C. edited and revised manuscript; J.-P.K.H., E.J.B., C.M.F., E.R.W., and R.C.C. approved final version of manuscript.

REFERENCES

1. **Arnold M**, **Goldschmitt M**, **Rigotti T.** Dealing with information overload: a comprehensive review. *Front Psychol* 14: 1122200, 2023. doi:10.3389/fpsyg.2023.1122200.
2. **Shahrzadi L**, **Mansouri A**, **Alavi M**, **Shabani A.** Causes, consequences, and strategies to deal with information overload: a scoping review. *Int J Inform Manage Data Insights* 4: 100261, 2024. doi:10.1016/j.jjimei.2024.100261.
3. **Pallivathukal RG**, **Kyaw Soe HH**, **Donald PM**, **Samson RS**, **Ismail AR.** ChatGPT for academic purposes: survey among undergraduate healthcare students in Malaysia. *Cureus* 16: e53032, 2024. doi:10.7759/cureus.53032.
4. **Abdellatif H**, **Al Mushaiqri M**, **Albalushi H**, **Al-Zaabi AA**, **Roychoudhury S**, **Das S.** Teaching, learning and assessing anatomy with artificial intelligence: the road to a better future. *Int J Environ Res Public Health* 19: 14209, 2022. doi:10.3390/ijerph192114209.
5. **Collins BR**, **Black EW**, **Rarey KE.** Introducing AnatomyGPT: a customized artificial intelligence application for anatomical sciences education. *Clin Anat* 37: 661–669, 2024. doi:10.1002/ca.24178.
6. **Favero TG.** Using artificial intelligence platforms to support student learning in physiology. *Adv Physiol Educ* 48: 193–199, 2024. doi:10.1152/advan.00213.2023.
7. **Gayed JM**, **Carlon MK**, **Oriola AM**, **Cross JS.** Exploring an AI-based writing Assistant's impact on English language learners. *Comput Educ Artif Intell* 3: 100055, 2022. doi:10.1016/j.caeai.2022.100055.
8. **Messeri L**, **Crockett MJ.** Artificial intelligence and illusions of understanding in scientific research. *Nature* 627: 49–58, 2024. doi:10.1038/s41586-024-07146-0.
9. **Desaire H**, **Chua AE**, **Isom M**, **Jarosova R**, **Hua D.** Distinguishing academic science writing from humans or ChatGPT with over 99% accuracy using off-the-shelf machine learning tools. *Cell Rep Phys Sci.* 4: 101426, 2023. doi:10.1016/j.xcrp.2023.101426.
10. **Perkins M**, **Roe J**, **Postma D**, **McGaughran J**, **Hickerson D.** Detection of GPT-4 generated text in higher education: combining academic judgement and software to identify generative AI tool misuse. *J Acad Ethics* 22: 89–113, 2024. doi:10.1007/s10805-023-09492-6.
11. **Habibzadeh F.** GPTZero performance in identifying artificial intelligence-generated medical texts: a preliminary study. *J Korean Med Sci* 38: e319, 2023. doi:10.3346/jkms.2023.38.e319.
12. **Walters WH.** The effectiveness of software designed to detect AI-generated writing: a comparison of 16 AI text detectors. *Open Inform Sci* 7: 20220158, 2023. doi:10.1515/opis-2022-0158.
13. **Malik AR**, **Pratiwi Y**, **Andajani K**, **Numertayasa IW**, **Suharti S**, **Darwis A**, **Marzuki.** Exploring artificial intelligence in academic essay: higher education student's perspective. *Int J Educ Res Open* 5: 100296, 2023. doi:10.1016/j.ijedro.2023.100296.
14. **Slominski T**, **Grindberg S**, **Momsen J.** Physiology is hard: a replication study of students' perceived learning difficulties. *Adv Physiol Educ* 43: 121–127, 2019. doi:10.1152/advan.00040.2018.
15. **Turki MA**, **Mohamud MS**, **Masuadi E**, **Altowejri MA**, **Farraj A**, **Schmidt HG.** The effect of using native versus nonnative language on the participation level of medical students during PBL tutorials. *Health Prof Educ* 6: 447–453, 2020. doi:10.1016/j.hpe.2020.11.001.
16. **Firetto CM**, **Starrett E**, **Montalbano AC**, **Yan Y**, **Penkrot TA**, **Kingsbury JS**, **Hyatt JP.** The impact of effective study strategy use in an introductory anatomy and physiology class. *Front Educ* 8: 1161772, 2023. doi:10.3389/feduc.2023.1161772.
17. **Yan L**, **Firetto CM**, **Starrett E**, **Kingsbury JS**, **Penkrot TA**, **Hyatt JP.** Exploring supports or incentives to promote undergraduate students' use of cooperative study groups. *Int J Educ Res Open* 4: 100252, 2023. doi:10.1016/j.ijedro.2023.100252.
18. **Chaka C.** Reviewing the performance of AI detection tools in differentiating between AI-generated and human-written texts: a literature and integrative hybrid review. *J Appl Learn Teach* 7: 115–126, 2024. doi:10.37074/jalt.2024.7.1.14.
19. **R Core Team.** R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, 2024. https://www.R-project.org.
20. **PositTeam.** R-Studio: Integrated Development Environment for R. Posit Software, 2024. http://www.posit.co.
21. **Agresti A.** *Categorical Data Analysis*. Wiley-Interscience Publications. 1990, p. 366–367.
22. **Chongsuvivatwong V.** epiDisplay: Epidemiological Data Display Package, R package version 3.5.0.2, 2022. https://CRAN.R-project.org/package=epiDisplay.
23. **Warnes GR**, **Bolker B**, **Lumley T**, **Johnson RC**, **Jain N**, **Schwartz M**, **Rogers J.** gmodels: various R Programming Tools for Model Fitting, R package version 2.19.1, 2024. https://CRAN.R-project.org/package=gmodels.
24. **Cannell B.** freqtables: Make Quick Descriptive Tables for Categorical Variables. R package version 0.1.1, 2022. https://CRAN.R-project.org/package=freqtables.
25. **McHugh ML.** Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)* 22: 276–282, 2012.
26. **Barton N.** Absence perception and the philosophy of zero. *Synthese* 197: 3823–3850, 2020. doi:10.1007/s11229-019-02220-x.
27. **Dalalah D**, **Dalalah OM.** The false-positives and false-negatives of generative AI detection tools in education and academic research: the case of ChatGPT. *Int J Manage Educ* 21: 100822, 2023. doi:10.1016/j.ijme.2023.100822.
28. **Waltzer T**, **Pilegard C**, **Heyman GD.** Can you spot the bot? Identifying AI-generated writing in college essays. *Int J Educ Integr* 20, 2024. doi:10.1007/s40979-024-00158-3.
29. **Liang W**, **Yuksekgonul M**, **Mao Y**, **Wu E**, **Zou J.** GPT detectors are biased against non-native English writers. *Patterns (NY)* 4: 100779, 2023. doi:10.1016/j.patter.2023.100779.