frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 07 June 2024
DOI 10.3389/feduc.2024.1374889

OPEN ACCESS

EDITED BY
Joana R. Casanova,
University of Minho, Portugal

REVIEWED BY
Keiichi Kobayashi,
Shizuoka University, Japan
Olga Vybornova,
Université Catholique de Louvain, Belgium

*CORRESPONDENCE
Timothy Paustian
paustian@wisc.edu

RECEIVED 22 January 2024
ACCEPTED 21 May 2024
PUBLISHED 07 June 2024

CITATION
Paustian T and Slinger B (2024) Students are using large language models and AI detectors can often detect their use.
Front. Educ. 9:1374889.
doi: 10.3389/feduc.2024.1374889

Students are using large language models and AI detectors can often detect their use

Timothy Paustian* and Betty Slinger

Department of Bacteriology, University of Wisconsin-Madison, Madison, WI, United States

Large language model (LLM) artificial intelligence (AI) has been in development for many years. Open AI thrust them into the spotlight in late 2022 when it released ChatGPT to the public. The wide availability of LLMs resulted in various reactions, from jubilance to fear. In academia, the potential for LLM abuse in written assignments was immediately recognized, with some instructors fearing they would have to eliminate this mode of evaluation. In this study, we seek to answer two questions. First, how are students using LLM in their college work? Second, how well do AI detectors function in the detection of AI-generated text? We organized 153 students from an introductory microbiology course to write essays on the regulation of the tryptophan operon. We then asked AI the same question and had the students try to disguise the answer. We also surveyed students about their use of LLMs. The survey found that 46.9% of students use LLM in their college work, but only 11.6% use it more than once a week. Students are unclear about what constitutes unethical use of LLMs. Unethical use of LLMs is a problem, with 39% of students admitting to using LLMs to answer assessments and 7% using them to write entire papers. We also tested their prose against five AI detectors. Overall, AI detectors could differentiate between human and AI-written text, identifying 88% correctly. Given the stakes, having a 12% error rate indicates we cannot rely on AI detectors alone to check LLM use, but they may still have value.

KEYWORDS

artificial intelligence, artificial intelligence detectors, plagiarism, cheating, large language model (LLM)

1 Introduction

Students have long used digital writing tools (spelling, style, and simple grammar checkers) to write assessments since their emergence in word processing programs in the late 1980s. These tools save the students time, help them learn writing skills, and result in a better final product. For years, autocorrect on phones has helped many a wayward finger but is sometimes the bane of anyone texting on their phone. More recently, writing assistants such as Grammarly, WordTune, and Perusall have helped students improve their writing, especially those where English is a second language. In most cases, these tools have been seen as helpful assistants to students, pointing out errors and allowing students to focus on core learning objectives (Perkins, 2023). All of these tools rely on some sort of artificial intelligence (AI).

In late 2022, the emergence of powerful large language model (LLM) artificial intelligence has scrambled the world of written communication. Some examples of large language models are BERT, GPT, Falcon, Ernie, and Palm, with more coming every month. Most LLMs are neural networks trained on large sets of textual data. A large proportion of the data used to

train LLMs is freely available on the Internet. LLMs then use their giant neural network to predict the next word of a sentence, which is repeated over and over to generate a complete response (Radford et al., 2019). The ability of these models to create human-like text and engage in conversations has generated significant interest in their abilities.

Educators are excited to explore these tools and determine how they could foster learning. LLMs can potentially change the focus in written tasks from mundane grammar to higher-level functions that engage the student with the material under study (Hess, 2023). Users can further enhance their prompts to the LLM to improve the response through conversation with the AI. Answers from the LLM can be phrased in plain language, making information easier to learn and helping people with communication disabilities (Hemsley et al., 2023). These models can serve as a tool for providing preliminary feedback to students and allow the instructor to focus on the content of their ideas, leaving the LLM to help the students with grammar and phrasing (Zawacki-Richter et al., 2019). Other forms of artificial intelligence (AI) can also identify at-risk students for intervention (Ouyang et al., 2022). The utilities of AI in higher education will expand as educators' experience increases.

Some educators also fear that these tools will short-circuit the learning process. Having students explain their understanding through written communication is one of the most effective forms of formative and summative assessment (Graham et al., 2015). We are all concerned our students will use LLM tools to create written assessments on their behalf, as nearly one-third of students report using Chat GPT (Intelligent, 2023). How can an instructor be sure that the ideas in a paper are those of the student and not AI?

The rapid emergence of LLMs, the apparent rapid adoption by some students, and the fevered discussion in society in general have universities playing catch up. Some universities have prohibited it outright, others have allowed it with restrictions, but most universities are hesitant to set policy without a larger time frame to assess its costs and benefits (Sullivan et al., 2023). Many of these policy decisions are difficult to make due to the newness of LLMs. Much of the opinions so far reported in the media and journals focus on the reactions of university staff and not student behavior. News media coverage of AI use in schools focuses on concerns about academic integrity and ways to discourage students from using LLMs in their academic work (Sullivan et al., 2023).

However, an important distinction here is how the students use LLMs. Most would define misconduct by a student as using an LLM, without attribution, to create the majority of the content of an assessment. When asked in the Intelligent survey (Intelligent, 2023), nearly 80% of students felt using an LLM was somewhat or definitely cheating, but the survey gave no details on how the students were using LLMs. While a few groups have surveyed students, it is still unclear how many students use LLMs and how exactly they use them. We need to know how students are using LLMs. They could be using it as a sophisticated form of information look-up to generate ideas for a writing assignment, to outline a paper to be written, to write the actual paper, to answer questions on a homework assignment, or to answer questions on an online exam. Most would agree the latter three uses would be academic misconduct. However, opinions vary on the first three uses of LLMs in assignments. The undetected use of AI can also have significant societal impacts, including mass propaganda through social media, news invented by LLMs, toxic spam to drive engagement, dishonest writing, fake product reviews, fake job applications, fake university application essays, or fake journal articles (Gillham, 2023). In addition, LLMs are known to hallucinate, making up facts or citations (Ye et al., 2023). Universities and other institutions must enact policies and procedures that ensure the transparent use of LLMs.

Nearly simultaneously with the rise of LLMs, detectors claiming to be able to detect content written by LLMs have emerged, including Open AI[1], Turniitin[2], GPTZero[3], ZeroGPT[4], Content at Scale[5], Winston[6], Originality.ai[7], and Packback[8]. These have met with mixed success; some assert their effectiveness, while others doubt their accuracy, but few independent studies of AI detectors have been undertaken. Liang et al. discovered that AI detectors would mistakenly flag non-native English speakers' writing as AI-generated (Liang et al., 2023). In some instances, faculty have unfairly accused students of cheating with AI based solely on the results of these detectors, resulting in significant controversies (Klee, 2023). Open AI eventually closed its detection tool due to its inability to differentiate between human and LLM-generated text. Other universities have decided to turn off the detection capabilities of some packages due to concerns about false positives (Coley, 2023).

Some studies have examined the ability of humans or AI-detection software to differentiate between human and AI-generated content. Small studies attempting to assess and train human graders to detect AI content have had limited success, with the graders identifying a significant amount of content incorrectly as human or AI-generated when the opposite was true (Clark et al., 2021; Gunser et al., 2021; Köbis and Mossink, 2021; Abd-Elaal et al., 2022). AI content detection tools have shown a better success rate. However, the occurrence of false positives and false negatives at too high a rate calls into question their usefulness (Elkhatat et al., 2023). A limitation of many of these studies is the small number of samples, especially human samples, tested. In addition, newer detectors are constantly appearing, as are tools that promise to avoid detection. An analysis using a larger group of students, focusing on a realistic assignment, would be useful. We also thought it would be interesting for students to work with a LLM and try to disguise the answer. Then, test AI detectors to see if they could correctly differentiate the writing samples.

In this study, we present an experiment carried out with the Fall 2023 cohort of introductory microbiology students at the University of Wisconsin-Madison. We asked students to write an essay of approximately 500 words explaining a topic in microbiology. They then created a prompt and submitted it to ChatGPT 3.5 or Google Bard to complete the same assignment. Finally, they attempted to disguise their answer to avoid AI detection. This process created a large dataset of 459 unique responses generated by individual students. We submitted all three essays from each student to five AI detectors: GPTZero, Wintson, Content at Scale, ZeroGPT, and Originality.ai. We chose these detectors because of their popularity, ability to

1 The Open AI detector was discontinued due to inaccuracy.
2 https://www.turnitin.com/
3 https://gptzero.me/
4 https://www.zerogpt.com/
5 https://contentatscale.ai/ai-content-detector/
6 https://app.gowinston.ai
7 https://app.originality.ai
8 https://www.packback.co/

be automated, and price. Students were also surveyed about their use of LLM in their academic studies.

The work found that 46.9% of students had at least explored LLMs. However, only 11.6% were using LLM on at least a weekly basis. The survey also showed that 7.2% had used LLMs to write an entire essay, and 39.2% had used it to answer questions on an exam or homework. We also found that the AI detectors GPTZero, ZeroGPT, and Orginatily.ai were successful at differentiating writing by students from that written by LLMs. Most students were unable to disguise their text and fool the detectors, but there were rare exceptions where the students were successful at disguising the text.

2 Materials and methods

2.1 Recruitment and class characteristics

Students enrolled in Microbiology 303 (The Biology of Prokaryotes) in the Fall of 2023 at UW-Madison were invited to participate in the study. Microbiology 303 is the introductory lecture for microbiology majors, and various majors in STEM fields also enroll in the course. Students were awarded 5 extra credit points to participate in the experiment but were given the option to leave at any time and still earn the extra credit. All students who attended the experiment decided to participate. Out of the class population of 224, 153 took part in the study. The racial breakdown of students was 74% white, 17.5% Asian/Pacific Islander, 4.5% Hispanic, 1.3% Black/African American, and 1.9% prefer not to say. Their year in school was 1.3% freshman, 13.6% sophomore, 47.4% junior, and 35.1% senior, with the rest being graduate students, a special student, and a non-degree-seeking student. The experimental design was submitted to the Madison Institutional Review Board (IRB), which determined that since the survey was anonymous and the focus of the research was the efficacy of the AI detectors, it did not constitute human subjects research (Submission ID No. 2023–1,548).

2.2 Administration of the survey

The survey (Supplementary material S1) consisted of students answering a question involving the tryptophan operon with or without the help of AI (as described above), and also completing several follow-up questions regarding their personal use of AI. The students signed up to participate in one of nine one-hour time slots from 29 November 2023 to 14 December 2023. Students took the survey in the presence of the experimenter to ensure a clear understanding of the survey and to prevent them from using LLMs in inappropriate places or plagiarizing answers from other sources. In the survey, they answered the following question:

> Explain the three levels of regulation of the tryptophan operon in *E. coli*. Make sure to include the proteins involved in each level and how they modulate the expression of the genes. Your answer should be about 500 words.

This topic was chosen because TP recently lectured on the regulation of the tryptophan operon of *Escherichia coli* and assessed them on the same material in an exam. Understanding bacterial regulation is a learning outcome in many microbiology courses, and the tryptophan operon is a common regulation paradigm. Students were allowed to use their notes to answer the question. Students then asked the same question of a LLM—either Google Bard (v. 2023.11.21 or 2023.12.06 versions)[9] or OpenAI's ChatGPT (v. 3.5)[10]. They could modify the prompt until they were satisfied with the answer. The students then added the unaltered AI answer to another part of the form. Next, the students attempted to modify the AI answer and disguise it in hopes of fooling an AI detector. The altered answer was put into a third part of the form. Finally, the students answered several questions about their use of AI. All responses were anonymous.

2.3 Testing of responses using AI detectors

All 153 survey responses were downloaded and saved as a CSV text file. We removed quotation marks (") in answers using Libre Office. This step prevents the quotation marks from confusing downstream scripts during processing. We developed a Python script to automate the process of submitting the students' answers to five AI detectors: GPTZero, Wintson, Content at Scale, ZeroGPT, and Originality.ai. The script took each answer and sent it to the detector website using an application programming interface or controlled the form interface at the website. The results of the AI check were then retrieved and recorded in a spreadsheet. The Python script is available in the Supplementary material S2. We recorded metrics for GPTZero, ZeroGPT, and Qriginality.ai as %AI (0–100) and for Winton as %Human. Content at Scale replied in one of three ways: *Passes as Human!*, *Hard to Tell!*, or *Reads like AI!*. In analyses, We converted the %Human value returned by Winston into %AI by subtracting %Human from 100.

2.4 Determining the accuracy of AI detectors

We determined the success of AI detectors as outlined by Gillham (2023). For comparisons, true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) were calculated. A TP was a human-written text that the AI detectors classified as ≤50% AI. A FP was a human-written text that was identified as >50%AI. A TN was an AI-written text that the AI detectors classified at >50% AI. Finally, a FN was an AI-written text that the AI detectors classified as ≤50% AI. We calculated the accuracy (a) of the detectors as follows:

$$a = \frac{TP + TN}{TP + FP + TN + FN}$$

Precision (p) as:

$$p = \frac{TP}{TP + FP}$$

9 https://bard.google.com/chat

10 https://platform.openai.com/apps

Recall (r), the true positive rate, as:

$$r = \frac{TP}{TP + FN}$$

Finally, we determined the overall performance of each detector by calculating an F1 score. This score takes into account both FP and FN.

$$F1 = \frac{p * r}{p + r}$$

The values for a, p, r, and F1 can range from 0 to 1.

2.5 Calculation of similarities

We calculated similarity measurements between the text written by students using the Python natural language toolkit (Bird et al., 2009). We used three methods: the cosine similarity, the Jaccard Similarity Index, and the Levenshtein distance. The cosine similarity measures the closeness of two sets of text that have been vectorized into multidimensional space. The Jaccard Similarity Index is the measurement of the similarity of two datasets. The texts to compare are transformed into sets, and the size of the intersection of the two sets is divided by their union. The Levenshtein distance indicates the number of changes required to transform one text into another. We measured similarities between human vs. AI, human vs. disguised, and AI vs. disguised.

2.6 Statistical analysis

We performed statistical analysis in R (R Development Core Team, 2022). We plotted histograms of each set of essays (human, AI, and disguised) for each detector. These plots suggested non-normalcy, and a Shapiro–Wilk normality test confirmed it. To test for statistical significance between the means, we performed a two-sample Wilcoxon's signed rank test between human vs. AI, human vs. disguised, and AI vs. disguised for each detector. We also used R to generate box plots of the detector results. In addition, we used R to create scatter plots comparing the Jaccard Similarity Index vs. AI-detection rates. The R commands to generate the plots are included in the Supplementary material S3.

2.7 Coding of student answers to survey questions with "other" as a choice

Three questions need to be coded for analysis. These were: Q11, "How have you used AI in your college work?," Q12, "Which of the following would you consider ethical uses of AI in your college work?," and Q13, "If you used AI in a way that you or your instructor might consider unethical, why did you do it?" We read and coded students' other responses, organizing them into categories. For Q11, we created 14 categories: Increase Understanding, Answer Questions on Homework, Answer Questions on Exam, Focus on Premise, Outline an Essay, Write an Essay, Editor/Grammar, Summarize Text, Study Guide Prompts, Writing outside of school (resumes, cover letters), Format/Find citations, and Find errors in code. For Q12, we created eight categories: Understanding Concepts, Premise/Title/Citations, Grammar, Outline Essay, Answer Homework Questions, Create Questions for Studying/Summarizing, None/Inaccurate, Write an Essay, and Answer questions on quiz or exam. For Q13, we created five categories: Lack of Time, After Large Effort, As Confirmation, Confusion with Writing/Reading, and Others are Using it.

2.8 Assessment of student, AI, and disguised answers

All 459 responses created by the students were graded using a rubric (Supplementary material S4). Before grading, the text was placed into a new spreadsheet, with each of the responses assigned a random code that hid the origin of the text (human, AI, and disguised) to prevent grading bias. A code sheet was also created that mapped each text response to its student ID and sample identity. After all samples were graded, the results were decoded using the code sheet, and the sample scores were compared.

3 Results

In total, 153 students explained the regulation of the tryptophan operon in about 500 words under direct supervision instead of having an online survey. These sessions allowed TP to clarify the directions and verify that students appropriately composed the three different pieces of writing. The goal of the writing exercise was to create authentic human writing, so students were allowed to use their notes, ask questions, or look up things online. The key was that they were to write their human response in their own words. Since the size of the groups was 30 students or less, we were able to verify that they wrote the human sections themselves. Students were allowed to work on the survey as long as they wanted, with most students finishing it in 30–60 min.

3.1 Students are using AI, but only occasionally

Students were surveyed about their LLM use. Over half (53.1%) have never used LLMs or experimented with them out of curiosity but decided they were not valuable to them. Approximately one-third (35.4%) use it a few times a month, 8.5% use it weekly, and 3.1% use it daily. The LLM used most often by students is ChatGPT (84.7%), with others mentioned being Chegg (6.1%), Bard (4.1%), Bing (3.1%), Snapchat AI (1%), and Quillbot (1%).

3.2 How are students using LLMs?

Students reported using AI in many ways, Figure 1. Many students (43 out of 153) are not using AI at all. In reading responses that students put in the other section of the question, these students gave two main reasons for not using AI: they "did not think it was helpful

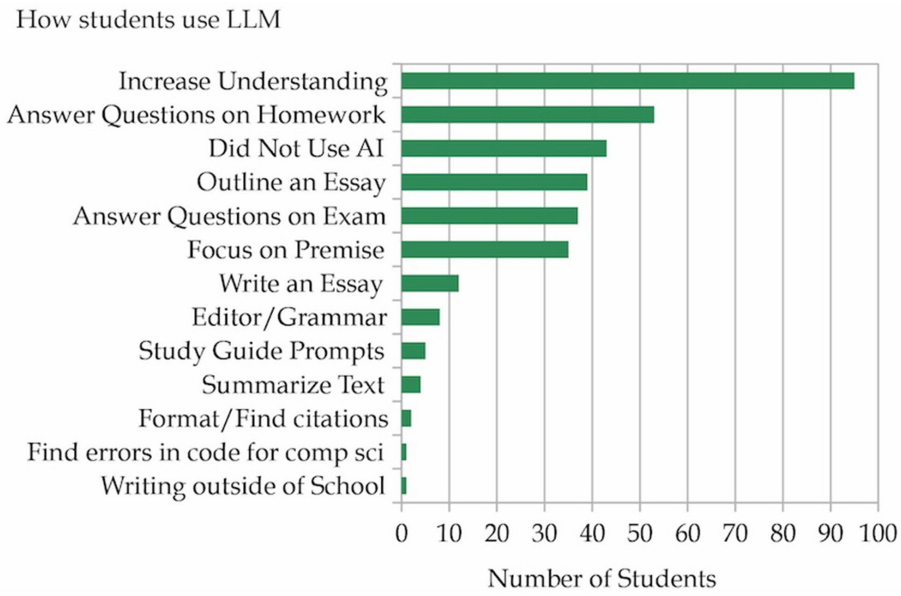

FIGURE 1
LLMs are used for many different tasks.

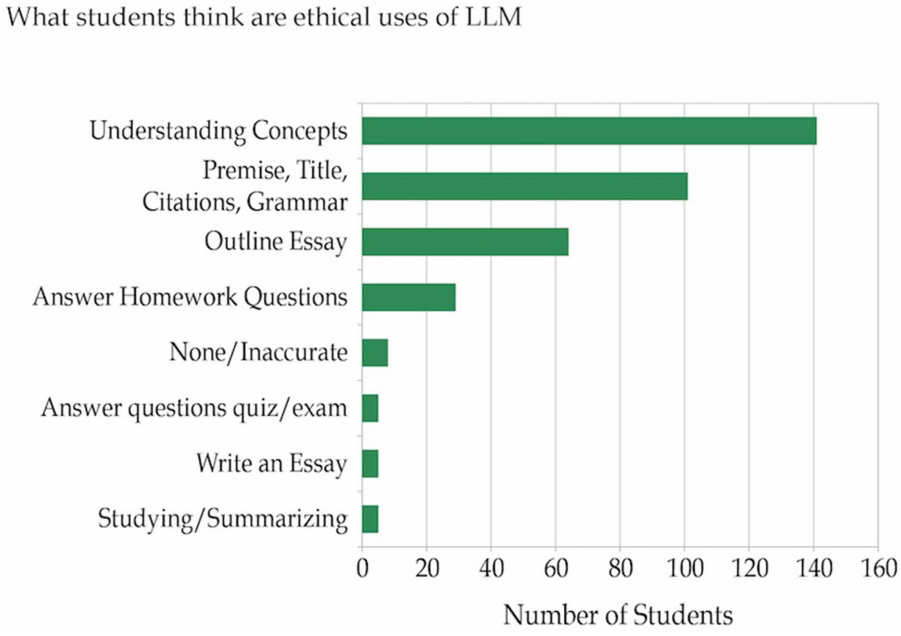

FIGURE 2
What students think are ethical uses of LLMs.

due to inaccuracies" or "they feared being accused of academic misconduct by their instructors." Students often used LLMs as a digital tutor, having the LLMs explain concepts they were trying to learn or summarize a text they were reading. They also reported using LLMs to offload some writing tasks, such as checking grammar and spelling, working on a premise, outlining a writing assignment, or even writing an entire article. Finally, students also used LLMs to answer questions on homework assignments or exams.

The survey also asked students what they would consider to be the ethical uses of LLM. Figure 2 shows the responses of 153 students. Most students (141) thought using LLM to understand concepts was ethical, and some more sophisticated users (5) used LLM to create questions or write summaries of topics under study. In writing, 101 thought finding premises for essays, improving grammar, or correctly formatting citations was ethical. While 64 felt more extensive uses, such as writing the outline, were appropriate, 5 even argued that having AI

write the first draft of the essay was acceptable. Some students (29) thought it was ethical to have LLM help them answer homework questions, and 5 thought having LLM answer exam and quiz questions was ethical. Finally, 5 students thought LLM had no ethical uses and were skeptical of its accuracy, especially in advanced classes.

For those students who did use the LLM in ways they thought others might consider unethical (54 students), the survey asked them why. Most often, students ran out of time or were stressed out (47.4%), or after trying to find an answer on their own and failing (19.3%), they would turn to the LLM for help. Others used AI as confirmation (15.8%) after they had answered a question. Some used AI when they felt their instructor did a poor job teaching concepts and they were confused about how to answer a question (8.8%). Finally, others justified using the LLM because they thought everyone else was using it (5.3%).

3.3 Students with a better understanding of the AI or the topic were no better at disguising their answers

Are students who have used AI better at disguising their answers than those who are naïve to AI? Students were divided into two groups: those who had used AI in the last 6 months (96 students) and those who had not (57 students). A comparison of their %AI scores of their human writing was identical, with both being marked with an average score of 16.1% AI with standard deviations of 31.3 for no AI use and 30.2 for the use of AI. The AI responses they submitted also had no difference, with AI scores of 98%. Finally, the deception text did show a small difference, with naïve students having an AI score of 81% while experienced students earning 74.3%. However, Wilcoxon's signed rank test showed a p-value of 0.31, indicating no significant difference between the means.

We also wondered if students who understood the topic would be able to better identify errors made by the AI and fix them and, in the process, also disguise the text from AI detectors. All 459 attempts to explain the regulation of the tryptophan operon were assessed using a rubric (Supplementary material S4). Before the assessment, the origin of the prose (human, AI, and disguised) was hidden to prevent grader bias. After grading, the scores were reassociated with each student for comparison. The scores were separated into two groups: students' human responses earning 100% of the points or higher (39 responses) and those earning <65% of the points (41 responses). Thus, we were comparing the students who answered the question well to those who scored the lowest on the question. The expectation was that students with better understanding might be able to better fool the detector. A comparison of disguised AI% scores for Originality.ai and ZeroGPT showed less than a 3.3% point difference between the higher scoring vs. lower scoring groups (Originality.ai: 73.3.0% vs. 72.1%) and (ZeroGPT: 52.4% vs. 55.7%).

A comparison of all the responses by students vs. AI showed that students were significantly better at answering the question overall, with human responses scoring 80% vs. 55% for the AI. Interestingly, when students tried to disguise their answers, they did not correct the mistakes the AI made, and there was very little improvement in the disguised score (57%). Many more students earned perfect scores (39) vs. the AI (6) or the disguised text (7).

3.4 Can AI detectors differentiate between human and AI-generated text?

We passed the text created by students through five AI detectors: Content at Scale, GPTZero, ZeroGPT, Winston, and Originality.ai. The AI detectors from GPTZero, ZeroGPT, and Orginality.ai reported values as %AI content. Winston's AI detector returned a %Human score. For comparisons, we converted Winston values to percent AI by subtracting the Winston %Human score from 100. Content at Scale returned three responses (Passes as Human!, Hard to Tell!, and Reads like AI!). While the AI-detector from Content at Scale had a recall rate of 0.89, its accuracy of (0.45) and precision of (0.47) indicate an anemic ability to identify AI-generated text. Due to the poor performance of Content at Scale, it will not be further analyzed in this study.

The other AI detectors had better success in correctly identifying human vs. AI-written text. Figure 3 shows a box plot of detector accuracy. In general, all four detectors could identify most of the generated content correctly. There was a clear distinction between human-generated text and AI-generated text. The students were somewhat successful in disguising their text, with the percentage AI value dropping 21% after being altered. However, the AI scores of the disguised text averaged 50 points higher than human samples. A two-sample Wilcoxon's signed rank test comparing human-*vs*-AI and human-*vs*-disguised for all detectors showed significant differences in means (Table 1). In the human and disguised samples, there were long tails on the box plots for some categories. The Winston detector failed badly when examining human text, flagging nearly half as false positives. GPTZero (24/152), Originality.ai (27/153), and ZeroGPT (15/153) had lower false positive rates. Thus, on average, for GPTZero, Originality, and ZeroGPT detectors, 14% of students would be detected using AI when they did not, and 5% of students who used AI would escape detection.

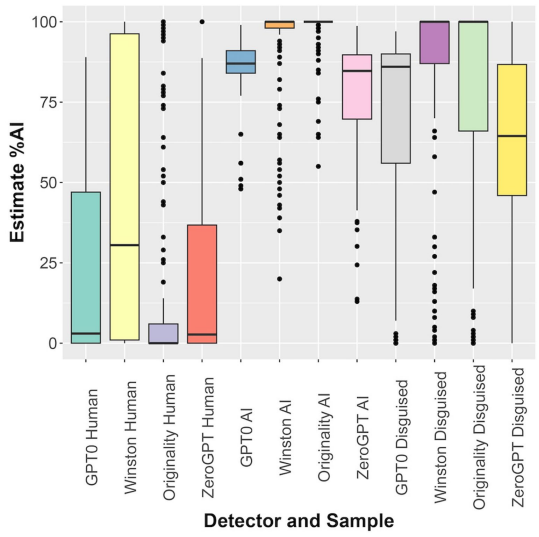

FIGURE 3
AI-detectors can differentiate human and AI-generated text authored by students. Box plots of estimated %AI of student writing for GPTZero, Winston, Orginality.ai, and ZeroGPT. Each detector was challenged with human, AI, and disguised text.

TABLE 1 AI benchmarks for detectors.

Detector	Wilcoxson (*p*)	Accuracy	Precision	Recall	F1
GPTZero (Human vs. AI)	< 2.2e-16	0.87	0.82	0.95	0.88
GPTZero (Human vs. Disguised)	< 2.2e-16	0.82	0.80	0.81	0.81
Winston (Human vs. AI)	< 2.2e-16	0.80	0.74	0.93	0.83
Winston (Human vs. Disguised)	5.053e-13	0.74	0.7	0.78	0.74
Originality.ai (Human vs. AI)	< 2.2e-16	0.91	0.85	1.0	0.92
Originality.ai (Human vs. Disguised)	< 2.2e-16	0.82	0.82	0.78	0.80
ZeroGPT (Human vs. AI)	< 2.2e-16	0.91	0.90	0.91	0.91
ZeroGPT (Human vs. Disguised)	< 2.2e-16	0.79	0.87	0.64	0.74

Several metrics are commonly used to assess the quality of AI detectors: accuracy (a), precision (p), recall (r), and F1 (see methods for details). Table 1 shows the results for each of the detectors when examining Human and AI-generated text. GPTZero, Originality.ai, ZeroGPT, and Winston did a comparable job, with Winston being a bit less precise (able to identify AI-generated text less accurately) than the others. If we exclude Winston's results, the F1 values of detectors were above 0.88.

Students' attempts to disguise the use of AI were somewhat successful (Table 1). All of the metrics dropped: recall (−20.6%), accuracy (−9.1%), precision (−3.7%), and F1 (−12.6%). Again, GPTZero, Orginality.ai, and ZeroGPT faired a little better than Winston, but the majority of the disguised text was still flagged as written by AI for all detectors.

An examination of the most successfully disguised text showed almost complete modification of the raw AI output (Figure 4). In panel A, the student reduced a response with an 89% AI score to 0 by substantial editing. (The non-highlighted text is the text the two responses have in common.) For those successful at disguising their AI answer, similarity metrics showed a large difference between the AI and disguised responses. In successfully disguised samples (<33% AI), the cosine similarity was 0.771 and the Jaccard Similarity Index was 0.501. In contrast, for those still detected as AI, the cosine similarity was 0.904 and the Jaccard Similarity Index was 0.69 between the AI-generated and disguised text. This lower similarity of the successfully disguised text shows that those who beat the detector successfully had done substantial editing. Panel B shows a disguised response that did not fool the detector. This failure to beat the detector is unsurprising since the student changed very little of the text. Examination of the several dozen or so responses that beat the detector showed a similar pattern. For students to beat the detectors, they had to rewrite the text substantially. It is possible that the students passed the initial AI response through AI summarizers such as AI Summarizer[11] or Quillbot[12] and did not edit the text themselves.

We plotted all four AI detector scores against the Jaccard Similarity Index between the AI-generated text and the disguised text, as shown in Figure 5. One would expect that as the students increased their modification of the text, the Jaccard Similarity Index would decrease, and the %AI score would also drop. The correlation roughly holds for GPTZero, ZeroGPT, and Originality.ai, with R values of 0.34, 0.36, and 0.38, respectively. The *p*-value for the Pearson correlation fit was significant for all three.

4 Discussion

This study organized a large group of microbiology students to answer a question about the regulation of the tryptophan operon. The same question was also posed to a LLM. Students were then tasked with trying to disguise the LLM answer to fool AI detectors. Students' creation of AI prompts and their attempts to fool the detector are the unique aspects of this study. We then assessed the ability of five AI detectors with the prose created by the students. Students also answered survey questions about their use of LLM.

There is a possibility that students were not truthful in the survey. However, we find it unlikely. The survey was anonymous, students were explicitly encouraged to be truthful at the beginning of each session, and there was no incentive for them to lie.

4.1 Half of the students had at least explored LLMs and some are using them in inappropriate ways

Of great interest to educators is how their students are using LLMs. This study found that about half of students are not using AI in their academic work. Another third used it only occasionally. We found that 11.6% of students routinely use AI in their college work. Other recent surveys found that approximately one-third of college students were using AI on written assignments (Intelligent, 2023). The use frequency of LLMs is lower than we expected (11.6% more than once a week). It appears right now that a majority of students are not

11 https://www.summarizer.org/

12 https://quillbot.com/

FIGURE 4
Successfully disguised text has significant changes from the AI response. The AI response is on the left and the disguised text is on the right. **(A)** A successfully disguised answer that dropped the AI score from 89 to 0%. **(B)** A failed attempt to disguise an answer that did not change in AI score. The non-highlighted text is the text the two responses have in common.

using LLMs in their studies. However, we anticipate that the proportion of students using AI in academic work will increase.

A major benefit of the current study was asking specific questions about students' behavior and thoughts on the use of LLMs. The survey by Intelligent (2023) asked one-third of students using LLM if they thought they were cheating, and over 75% thought they were but did it anyway. In this study, we asked students about their behaviors in more detail. Much of the use of LLMs (66%) was to understand concepts, double-check answers, help focus on a premise, summarize text, or perform grammar/editor functions. These uses would be acceptable to many instructors. It was concerning how many students (39%) were using LLMs to answer questions on homework or out-of-class exams. While other questions in the survey indicated that some of this work was double-checking answers, too many students are using it to cheat. The poor performance of LLMs in answering a basic question in bacterial regulation should serve as a

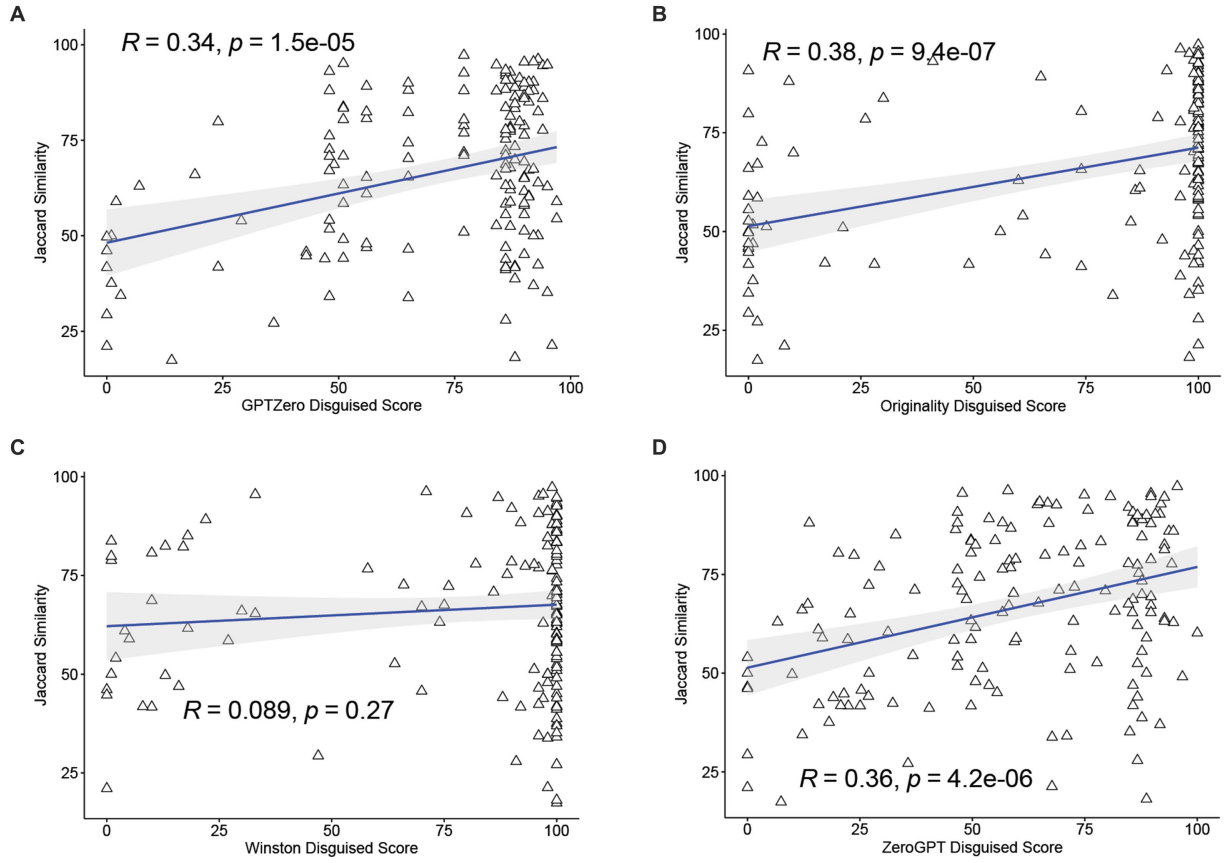

FIGURE 5
The more human editing that is done to disguise a text, the lower the AI score. A comparison of AI score vs. Jaccard Similarity Index between AI-generated text and disguised text. GPTZero **(A)**, Originality **(B)**, Winton **(C)**, ZeroGPT **(D)**. For three of the four detectors **(A,B,D)**, as the Jaccard Similarity Index decreases, so does the AI score.

cautionary tale to students. Instructors may have to redesign written, out-of-class quizzes and exams because, clearly, many students will use LLMs to complete them. A small minority (7%) of students reported using AI to write entire papers.

Students are just beginning to explore the uses of LLMs for their studies, and they have some sense of ethical uses of the technology. However, there is confusion about the boundaries, with some students believing that using AI to create outlines or even write first drafts of their academic work is acceptable. Many instructors would probably disagree. Students need to know the acceptable uses of AI in their classroom. It is imperative that instructors, administrators, and colleges set clear guidelines for students *and* instructors. Due to the rapid development and deployment of LLMs, this is an urgent priority, and higher education can no longer take a wait-and-see attitude.

When asked why students used LLMs in ways they would consider unethical, the number one answer was a lack of time. Students also listed a failure to find the answer on their own, confusion about what the question was asking, and because they thought others were using LLMs to cheat. These responses are in line with recent work that explains cheating by situational motivations (Waltzer and Dahl, 2023). The perception that using AI in these ways is not cheating, along with factors such as the need for a good grade, may override students' motivation to be honest.

The survey population was restricted to 153 students taking a microbiology course. Larger surveys, with greater diversity in race, class, and type of college, would be valuable.

4.2 Familiarity with using an AI does not correlate with avoiding detection

Understanding of the topic or familiarity with AI tools did not impact the ability of the students to evade detection by the AI detectors. It seems as if knowing how to use an AI or being skillful at writing effective prompts did not correlate with an ability to avoid detection. Avoiding detection seems to be a separate skill set, one that is not honed by using AI. Avoiding detection requires significant rewriting by the student, to the point that just writing the assignment themselves would probably be less effort.

The grading of the answer to tryptophan regulation indicated that many students on their own understood the regulation of the operon and could explain it well. However, the AI responses earned close to a failing grade. General LLMs can create responses that, on the surface, seem accurate. However, when the LLM was asked to explain what we would consider basic regulation in bacteria, it was not up to the task.

In reading through all the AI responses, there were wildly inaccurate explanations stated in confident language. One common error was the inclusion of the *trp* RNA-binding attenuation protein in discussing the regulation of tryptophan in *Escherichia coli*. This protein is not found in *E. coli* but instead is present in *Bacillus subtilis* tryptophan regulation. A second common error was the insertion of catabolite repression, a system that regulates the expression of carbohydrate degradation genes in *E. coli* but is not involved in tryptophan biosynthesis. It appears that AI struggles significantly when it is asked to produce text in an area of specialized knowledge. The limited ability of AI to answer more specific questions is unsurprising since LLMs are trained on publicly available text. There is too large of a probability that training text will contain inaccurate information, thus confusing the LLM. A common misconception for students is to mix up the behavior of the secondary structure in attenuation and describe the opposite result. In other words, they will think low concentrations of tryptophan lead to the formation of the rho-independent terminator that stops transcription. We often found this error in LLM explanations, suggesting it may have picked up this misconception from errant pages describing the process. LLM explanations also frequently skipped the post-translation mechanism of feedback inhibition. This reflects a common misunderstanding of bacterial regulation, where the focus is solely on gene expression and later regulation points are ignored. The LLM did get right the simpler facets of tryptophan regulation, namely the behavior of the *trp* repressor.

4.3 AI detectors work, but not well enough to stand alone

Four of the five AI detectors tested were able to identify AI-generated text. The Content at Scale detector failed to identify over half of the AI-written text as AI. However, the four other detectors could differentiate human vs. AI-generated text. A comparison of AI-score distributions between human vs. AI and human vs. disguised text showed that GPTZero, Winston, ZeroGPT, and Originality.ai all showed highly significant differences between the means. All four detectors had false negative rates below 9%, with Originality.ai having no false negatives. The accuracy, precision, recall, and F1 were all above 0.7 in all cases.

Nevertheless, AI detectors need to be nearly perfect for them to be trusted to take on the role of policing student writing. The sticking point is false positives. Instructors do not want to accuse students of academic dishonesty unless they are certain it exists and any error, where the detectors flag human writing as AI, is problematic. Unfortunately, GPTZero, Winston, ZeroGPT, and Originality.ai too often identified human writing as written by AI, with rates of 15.6, 45.8, 9.8, and 17.6%, respectively. The high rate of false positives from the Winston detector makes it unusable to monitor students. One method of reducing false positives is to use two detectors in combination. When a piece of human writing was passed through both GPTZero and Originality.ai and only counted as suspicious if both detectors flagged it, the false positive rate dropped to 5.2%. Still, this is too high to be relied on alone and creates more false negatives.

It is clear from this study that AI detectors cannot be relied upon as the only metric to determine the use of AI by students. However, it is also true that AI detectors are generally able to flag the use of AI. Instructors may be able to use AI detectors as one tool to incentivize students to do their own work.

This study provides a snapshot of students' current use of LLMs and the capability of AI detectors. While students are beginning to use LLMs, their use is not universal. Of those who do use LLMs, many are using them ethically, but too many use them in ways their instructors would probably consider inappropriate. It is possible to use AI detectors as one component of a comprehensive policy to encourage ethical uses of LLMs. As these technologies develop, the landscape is sure to shift. Instructors and institutions must stay current on the latest technologies and create supportive environments where students understand the responsible use of LLMs.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The experimental design was submitted to the Madison IRB which determined that since the survey was anonymous and the focus of the research was the efficacy of the AI detectors, it did not constitute human subjects research. (Submission ID No. 2023-1548). Written informed consent for participation in the study was not required from the participants in accordance with the local legislation and institutional requirements.

Author contributions

TP: Conceptualization, Data curation, Formal analysis, Methodology, Validation, Writing – original draft, Writing – review & editing. BS: Conceptualization, Formal analysis, Methodology, Validation, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

We would like to thank the students of the Fall class of 2023 of Microbiology 303. Their enthusiasm and willingness to participate in the study made it possible. We would also like to thank Dr. Michelle Rondon and Dr. Melissa Christopherson for reviewing this manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2024.1374889/full#supplementary-material

References

Abd-Elaal, E. S., Gamage, S. H. P. W., and Mills, J. E. (2022). Assisting academics to identify computer generated writing. *Eur. J. Eng. Educ.* 47, 725–745. doi: 10.1080/03043797.2022.2046709

Bird, S., Klein, E., and Loper, E. (2009), *Natural language processing with Python: Analyzing text with the natural.*

Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., and Smith, N.A. (2021), "All That's 'human' is not gold: evaluating human evaluation of generated text", ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference, Association for Computational Linguistics (ACL), pp. 7282–7296

Coley, M. (2023), "Guidance on AI detection and why We're disabling Turnitin's AI detector | Brightspace support | Vanderbilt University". Available at:https://www.vanderbilt.edu/brightspace/2023/08/16/guidance-on-ai-detection-and-why-were-disabling-turnitins-ai-detector/ Accessed on December, 12 2023.

Elkhatat, A. M., Elsaid, K., and Almeer, S. (2023). Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *Int. J. Educ. Integr.* 19:17. doi: 10.1007/s40979-023-00140-5

Gillham, J. (2023), "AI content detector accuracy review + open source dataset and research tool – originality.AI". Available at:https://originality.ai/blog/ai-content-detection-accuracy Accessed on December, 27 2023.

Graham, S., Hebert, M., and Harris, K. R. (2015). Formative assessment and writing. *Elem. Sch. J.* 115, 523–547. doi: 10.1086/681947

Gunser, V. E., Gottschling, S., Brucker, B., Richter, S., and Gerjets, P. (2021). Can users distinguish narrative texts written by an artificial intelligence writing tool from purely human text? *Commun. Comp. Infor. Sci.* 1419, 520–527. doi: 10.1007/978-3-030-78635-9_67

Intelligent. (2023), "Nearly 1 in 3 college students have used ChatGPT on written assignments - Intelligent", Intelligent. Available at:https://www.intelligent.com/nearly-1-in-3-college-students-have-used-chatgpt-on-written-assignments/ Accessed on December, 12 2023.

Klee, M. (2023), "Texas a&M professor wrongly accuses class of cheating with ChatGPT", rolling stone. Available at:https://www.rollingstone.com/culture/culture-features/texas-am-chatgpt-ai-professor-flunks-students-false-claims-1234736601/ Accessed on December, 12 2023).

Köbis, N., and Mossink, L. D. (2021). Artificial intelligence versus Maya Angelou: experimental evidence that people cannot differentiate AI-generated from human-written poetry. *Comput. Hum. Behav.* 114:106553. doi: 10.1016/J.CHB.2020.106553

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., and Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns* 4:100779. doi: 10.1016/j.patter.2023.100779

R Development Core Team. (2022), *R: a language and environment for statistical computing*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). *Language models are unsupervised multitask learners*: OpenAI Blog.

Waltzer, T., and Dahl, A. (2023). Why do students cheat? Perceptions, evaluations, and motivations. *Ethics Behav.* 33, 130–150. doi: 10.1080/10508422.2022.2026775

Ye, H., Liu, T., Zhang, A., Hua, W., Jia, W., and Lab, Z. (2023), "*Cognitive mirage: A review of hallucinations in large language models*".