International Journal for Educational Integrity

ORIGINAL ARTICLE Open Access

Can you spot the bot? Identifying AI-generated writing in college essays

Tal Waltzer[1*], Celeste Pilegard[1] and Gail D. Heyman[1]

*Correspondence:
Tal Waltzer
twaltzer@ucsd.edu
[1]Department of Psychology, University of California San Diego, 9500 Gilman Drive, La Jolla, San Diego, CA 92093-0109, USA

Abstract

The release of ChatGPT in 2022 has generated extensive speculation about how Artificial Intelligence (AI) will impact the capacity of institutions for higher learning to achieve their central missions of promoting learning and certifying knowledge. Our main questions were whether people could identify AI-generated text and whether factors such as expertise or confidence would predict this ability. The present research provides empirical data to inform these speculations through an assessment given to a convenience sample of 140 college instructors and 145 college students (Study 1) as well as to ChatGPT itself (Study 2). The assessment was administered in an online survey and included an AI Identification Test which presented pairs of essays: In each case, one was written by a college student during an in-class exam and the other was generated by ChatGPT. Analyses with binomial tests and linear modeling suggested that the AI Identification Test was challenging: On average, instructors were able to guess which one was written by ChatGPT only 70% of the time (compared to 60% for students and 63% for ChatGPT). Neither experience with ChatGPT nor content expertise improved performance. Even people who were confident in their abilities struggled with the test. ChatGPT responses reflected much more confidence than human participants despite performing just as poorly. ChatGPT responses on an AI Attitude Assessment measure were similar to those reported by instructors and students except that ChatGPT rated several AI uses more favorably and indicated substantially more optimism about the positive educational benefits of AI. The findings highlight challenges for scholars and practitioners to consider as they navigate the integration of AI in education.

Keywords ChatGPT, Artificial intelligence, Cheating, Academic integrity, Confidence, Higher education

Introduction

Artificial intelligence (AI) is becoming ubiquitous in daily life. It has the potential to help solve many of society's most complex and important problems, such as improving the detection, diagnosis, and treatment of chronic disease (Jiang et al. 2017), and informing public policy regarding climate change (Biswas 2023). However, AI also comes with potential pitfalls, such as threatening widely-held values like fairness and the right to privacy (Borenstein and Howard 2021; Weidinger et al. 2021; Zhuo et al. 2023). Although

the specific ways in which the promises and pitfalls of AI will play out remain to be seen, it is clear that AI will change human societies in significant ways.

In late November of 2022, the generative large-language model ChatGPT (GPT-3, Brown et al. 2020) was released to the public. It soon became clear that talk about the consequences of AI was much more than futuristic speculation, and that we are now watching its consequences unfold before our eyes in real time. This is not only because the technology is now easily accessible to the general public, but also because of its advanced capacities, including a sophisticated ability to use context to generate appropriate responses to a wide range of prompts (Devlin et al. 2018; Gilson et al. 2022; Susnjak 2022; Vaswani et al. 2017).

How AI-generated content poses challenges for educational assessment

Since AI technologies like ChatGPT can flexibly produce human-like content, this raises the possibility that students may use the technology to complete their academic work for them, and that instructors may not be able to tell when their students turn in such AI-assisted work. This possibility has led some people to argue that we may be seeing the end of essay assignments in education (Mitchell 2022; Stokel-Walker 2022). Even some advocates of AI in the classroom have expressed concerns about its potential for undermining academic integrity (Cotton et al. 2023; Eke 2023). For example, as Kasneci et al. (2023) noted, the technology might "amplify laziness and counteract the learners' interest to conduct their own investigations and come to their own conclusions or solutions" (p. 5). In response to these concerns, some educational institutions have already tried to ban ChatGPT (Johnson, 2023; Rosenzweig-Ziff 2023; Schulten, 2023).

These discussions are founded on extensive scholarship on academic integrity, which is fundamental to ethics in higher education (Bertram Gallant 2011; Bretag 2016; Rettinger and Bertram Gallant 2022). Challenges to academic integrity are not new: Students have long found and used tools to circumvent the work their teachers assign to them, and research on these behaviors spans nearly a century (Cizek 1999; Hartshorne and May 1928; McCabe et al. 2012). One recent example is contract cheating, where students pay other people to do their schoolwork for them, such as writing an essay (Bretag et al. 2019; Curtis and Clare 2017). While very few students (less than 5% by most estimates) tend to use contract cheating, AI has the potential to make cheating more accessible and affordable and it raises many new questions about the relationship between technology, academic integrity, and ethics in education (Cotton et al. 2023; Eke 2023; Susnjak 2022).

To date, there is very little empirical evidence to inform debates about the likely impact of ChatGPT on education or to inform what best practices might look like regarding use of the technology (Dwivedi et al. 2023; Lo 2023). The primary goal of the present research is to provide such evidence with reference to college-essay writing. One critical question is whether college students can pass off work generated by ChatGPT as their own. If so, large numbers of students may simply paste in ChatGPT responses to essays they are asked to write without the kind of active engagement with the material that leads to deep learning (Chi and Wylie 2014). This problem is likely to be exacerbated when students brag about doing this and earning high scores, which can encourage other students to follow suit. Indeed, this kind of bragging motivated the present work (when the last author learned about a college student bragging about using ChatGPT to write all of her final papers in her college classes and getting A's on all of them).

In support of the possibility that instructors may have trouble identifying ChatGPT-generated test, some previous research suggests that ChatGPT is capable of successfully generating college- or graduate-school level writing. Yeadon et al. (2023) used AI to generate responses to essays based on a set of prompts used in a physics module that was in current use and asked graders to evaluate the responses. An example prompt they used was: "How did natural philosophers' understanding of electricity change during the 18th and 19th centuries?" The researchers found that the AI-generated responses earned scores comparable to most students taking the module and concluded that current AI large-language models pose "a significant threat to the fidelity of short-form essays as an assessment method in Physics courses." Terwiesch (2023) found that ChatGPT scored at a B or B- level on the final exam of Operations Management in an MBA program, and Katz et al. (2023) found that ChatGPT has the necessary legal knowledge, reading comprehension, and writing ability to pass the Bar exam in nearly all jurisdictions in the United States. This evidence makes it very clear that ChatGPT can generate well-written content in response to a wide range of prompts.

Distinguishing AI-generated from human-generated work

What is still not clear is how good instructors are at distinguishing between ChatGPT-generated writing and writing generated by students at the college level given that it is at least possible that ChatGPT-generated writing could be both high quality and be distinctly different than anything people generally write (e.g., because ChatGPT-generated writing has particular features). To our knowledge, this question has not yet been addressed, but a few prior studies have examined related questions. In the first such study, Gunser et al. (2021) used writing generated by a ChatGPT predecessor, GPT-2 (see Radford et al. 2019). They tested nine participants with a professional background in literature. These participants both generated content (i.e., wrote continuations after receiving the first few lines of unfamiliar poems or stories), and determined how other writing was generated. Gunser et al. (2021) found that misclassifications were relatively common. For example, in 18% of cases participants judged AI-assisted writing to be human-generated. This suggests that even AI technology that is substantially less advanced than ChatGPT is capable of generating writing that is hard to distinguish from human writing.

Köbis and Mossink (2021) also examined participants' ability to distinguish between poetry written by GPT-2 and humans. Their participants were given pairs of poems. They were told that one poem in each pair was written by a human and the other was written by GPT-2, and they were asked to determine which was which. In one of their studies, the human-written poems were written by professional poets. The researchers generated multiple poems in response to prompts, and they found that when the comparison GPT-2 poems were ones they selected as the best among the set generated by the AI, participants could not distinguish between the GPT-2 and human writing. However, when researchers randomly selected poems generated by GPT-2, participants were better than chance at detecting which ones were generated by the AI.

In a third relevant study, Waltzer et al. (2023a) tested high school teachers and students. All participants were presented with pairs of English essays, such as one on why literature matters. In each case one essay was written by a high school student and the other was generated by ChatGPT, and participants were asked which essay in each pair

had been generated by ChatGPT. Waltzer et al. (2023a) found that teachers only got it right 70% of the time, and that students' performance was even worse (62%). They also found that well-written essays were harder to distinguish from those generated by ChatGPT than poorly written ones. However, it is unclear the extent to which these findings are specific to the high school context. It should also be noted that there were no clear right or wrong answers in the types of essays used in Waltzer et al. (2023a), so the results may not generalize to essays that ask for factual information based on specific class content.

AI detection skills, attitudes, and perceptions

If college instructors find it challenging to distinguish between writing generated by ChatGPT and college students, it raises the question of what factors might be correlated with the ability to perform this discrimination. One possible correlate is experience with ChatGPT, which may allow people to recognize patterns in the writing style it generates, such as a tendency to formally summarize previous content. Content-relevant knowledge is another possible predictor. Individuals with such knowledge will presumably be better at spotting errors in answers, and it is plausible that instructors know that AI tools are likely to get content of introductory-level college courses correct and assume that essays that contain errors are written by students.

Another possible predictor is confidence about one's ability to discriminate on the task or on particular items of the task (Erickson and Heit 2015; Fischer & Budesco, 2005; Wixted and Wells 2017). In other words, are AI discriminations made with a high degree of confidence more likely to be accurate than low-confidence discriminations? In some cases, confidence judgments are a good predictor of accuracy, such as on many perceptual decision tasks (e.g., detecting contrast between light and dark bars, Fleming et al. 2010). However, in other cases correlations between confidence and accuracy are small or non-existent, such as on some deductive reasoning tasks (e.g., Shynkaruk and Thompson 2006). Links to confidence can also depend on how confidence is measured: Gigerenzer et al. (1991) found overconfidence on individual items, but good calibration when participants were asked how many items they got right after seeing many items.

In addition to the importance of gathering empirical data on the extent to which instructors can distinguish ChatGPT from college student writing, it is important to examine how college instructors and students perceive AI in education given that such attitudes may affect behavior (Al Darayseh 2023; Chocarro et al. 2023; Joo et al. 2018; Tlili et al. 2023). For example, instructors may only try to develop precautions to prevent AI cheating if they view this as a significant concern. Similarly, students' confusion about what counts as cheating can play an important role in their cheating decisions (Waltzer and Dahl 2023; Waltzer et al. 2023b).

The present research

In the present research we developed an assessment that we gave to college instructors and students (Study 1) and ChatGPT itself (Study 2). The central feature of the assessment was an *AI Identification Test*, which included 6 pairs of essays. In each case (as was indicated in the instructions), one essay in each pair was generated by ChatGPT and the other was written by college students. The task was to determine which essay was written by the chatbot. The essay pairs were drawn from larger pools of essays of each type.

The student essays were written by students as part of a graded exam in a psychology class, and the ChatGPT essays were generated in response to the same essay prompts. Of interest was overall performance and to assess potential correlates of performance. Performance of college instructors was of particular interest because they are the ones typically responsible for grading, but performance of students and ChatGPT were also of interest for comparison. ChatGPT was also of interest given anecdotal evidence that college instructors are asking ChatGPT to tell them whether pieces of work were AI-generated. For example, the academic integrity office at one major university sent out an announcement asking instructors not to report students for cheating if their evidence was solely based on using ChatGPT to detect AI-generated writing (UCSD Academic Integrity Office, 2023).

We also administered an *AI Attitude Assessment* (Waltzer et al. 2023a), which included questions about overall levels of optimism and pessimism about the use of AI in education, and the appropriateness of specific uses of AI in academic settings, such as a student submitting an edited version of a ChatGPT-generated essay for a writing assignment.

Study 1: College instructors and students

Method

Participants were given an online assessment that included an *AI Identification Test*, an *AI Attitude Assessment*, and some demographic questions. The AI Identification Test was developed for the present research, as described below (see Materials and Procedure). The test involved presenting six pairs of essays, with the instructions to try to identify which one was written by ChatGPT in each case. Participants also rated their confidence before the task and after responding to each item, and reported how many they thought they got right at the end. The AI Attitude Assessment was drawn from Waltzer et al. (2023a) to assess participants' views of the use of AI in education.

Participants

For the testing phase of the project, we recruited 140 instructors who had taught or worked as a teaching assistant for classes at the college level (69 of them taught psychology and 63 taught other subjects such as philosophy, computer science, and history). We recruited instructors through personal connections and snowball sampling. Most of the instructors were women (59%), white (60%), and native English speakers (67%), and most of them taught at colleges in the United States (91%). We also recruited 145 undergraduate students (M_{age} = 20.90 years, 80% women, 52% Asian, 63% native English speakers) from a subject recruitment system in the psychology department at a large research university in the United States. All data collection took place between 3/15/2023 and 4/15/2023 and followed our pre-registration plan (https://aspredicted.org/mk3a2.pdf).

Materials and procedure

Developing the AI identification test To create the stimuli for the AI Identification Test, we first generated two prompts for the essays (Table 1). We chose these prompts in collaboration with an instructor to reflect real student assignments for a college psychology class.

Table 1 *Prompts for Generating Essays*

	Shorthand	Full Prompt
1	Phonemic Awareness	Write a definition, example, and explanation for the following term: phonemic awareness Definition: Write a concise statement of what the term means. Example: Describe a concrete and specific instance of the term. The example can be made up. Explanation: Provide a theoretical implication of the term or state an important scientific fact about it. Your complete answer should be 3–6 sentences long. Do not assume that the grader already knows the answer. Use complete English sentences.
2	Studying Advice	How should college students study? Name and describe two relevant concepts from the science of learning, instruction, and/or assessment. For each concept, apply your understanding to describe a concrete piece of advice for college students to study effectively. Your complete answer should be 4–8 sentences long. Do not assume that the grader already knows the answer. Use complete English sentences.

Fifty undergraduate students hand-wrote both essays as part of a proctored exam in their psychology class on 1/30/2023. Research assistants transcribed the essays and removed essays from the pool that were not written in third-person or did not include the correct number of sentences. Three additional essays were excluded for being illegible, and another one was excluded for mentioning a specific location on campus. This led to 15 exclusions for the Phonemic Awareness prompt and 25 exclusions for the Studying Advice prompt. After applying these exclusions, we randomly selected 25 essays for each prompt to generate the 6 pairs given to each participant. To prepare the texts for use as stimuli, research assistants then used a word processor to correct obvious errors that could be corrected without major rewriting (e.g., punctuation, spelling, and capitalization).

All student essays were graded according to the class rubric on a scale from 0 to 10 by two individuals on the teaching team of the class: the course's primary instructor and a graduate student teaching assistant. Grades were averaged together to create one combined grade for each essay (mean: 7.93, SD: 2.29, range: 2–10). Two of the authors also scored the student essays for writing quality on a scale from 0 to 100, including clarity, conciseness, and coherence (combined score mean: 82.83, *SD*: 7.53, range: 65–98). Materials for the study, including detailed scoring rubrics, are available at https://osf.io/2c54a/.

The ChatGPT stimuli were prepared by entering the same prompts into ChatGPT (https://chat.openai.com/) between 1/23/2023 and 1/25/2023, and re-generating the responses until there were 25 different essays for each prompt.

Testing Phase In the participant testing phase, college instructors and students took the assessment, which lasted approximately 10 min. All participants began by indicating the name of their school and whether they were an instructor or a student, how familiar they were with ChatGPT ("Please rate how much experience you have with using ChatGPT"), and how confident they were that they would be able to distinguish between writing generated by ChatGPT and by college students. Then they were told they would get to see how well they score at the end, and they began the AI Identification Test.

The AI Identification Test consisted of six pairs of essays: three Phonemic Awareness pairs, and three Studying Advice pairs, in counterbalanced order. Each pair included one text generated by ChatGPT and one text generated by a college student, both drawn

randomly from their respective pools of 25 possible essays. No essays were repeated for the same participant. Figure 1 illustrates what a text pair looked like in the survey.

For each pair, participants selected the essay they thought was generated by ChatGPT and indicated how confident they were about their choice (slider from 0 = "not at all confident" to 100 = "extremely confident"). After all six pairs, participants estimated how well they did ("How many of the text pairs do you think you answered correctly?").

After completing the AI Identification task, participants completed the AI Attitude Assessment concerning their views of ChatGPT in educational contexts (see Waltzer et al. 2023a). On this assessment, participants first estimated what percent of college students in the United States would ask ChatGPT to write an essay for them and submit it. Next, they rated their concerns ("How concerned are you about ChatGPT having negative effects on education?") and optimism ("How optimistic are you about ChatGPT having positive benefits for education?") about the technology on a scale from 0 ("not at all") to 100 ("extremely"). On the final part of the AI Attitude Assessment, they evaluated five different possible uses of ChatGPT in education (such as submitting an essay after asking ChatGPT to improve the vocabulary) on a scale from −10 ("really bad") to +10 ("really good").

Participants also rated the extent to which they already knew the subject matter (i.e., cognitive psychology and the science of learning), and were given optional open-ended text boxes to share any experiences from their classes or suggestions for instructors related to the use of ChatGPT, or to comment on any of the questions in the Attitude Assessment. Instructors were also asked whether they had ever taught a psychology class and to describe their teaching experience. At the end, all participants reported demographic information (e.g., age, gender). All prompts are available in the online supplementary materials (https://osf.io/2c54a/).

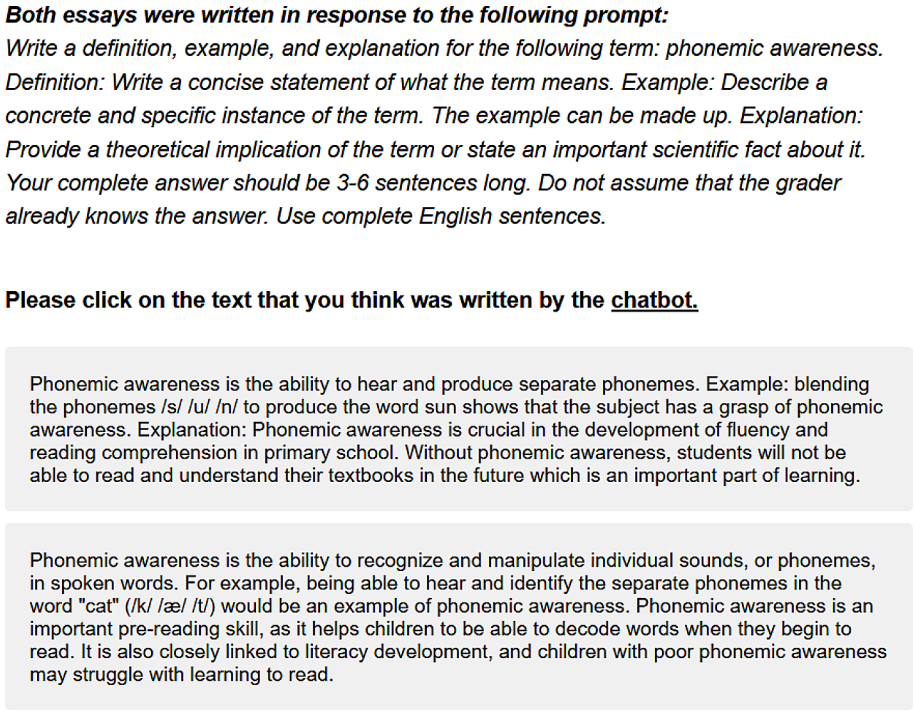

Fig. 1 Example pair of essays for the Phonemic Awareness prompt. Top: student essay. Bottom: ChatGPT essay

Data Analysis

We descriptively summarized variables of interest (e.g., overall accuracy on the Identification Test). We used inferential tests to predict Identification Test accuracy from group (instructor or student), confidence, subject expertise, and familiarity with ChatGPT. We also predicted responses to the AI Attitude Assessment as a function of group (instructor or student). All data analysis was done using R Statistical Software (v4.3.2; R Core Team 2021).

Key hypotheses were tested using Welch's two-sample t-tests for group comparisons, linear regression models with F-tests for other predictors of accuracy, and Generalized Linear Mixed Models (GLMMs, Hox 2010) with likelihood ratio tests for within-subjects trial-by-trial analyses. GLMMs used random intercepts for participants and predicted trial performance (correct or incorrect) using trial confidence and essay quality as fixed effects.

Results

Overall performance on AI identification test

Instructors correctly identified which essay was written by the chatbot 70% of the time, which was above chance (chance: 50%, binomial test: $p < .001$, 95% CI: [66%, 73%]). Students also performed above chance, with an average score of 60% (binomial test: $p < .001$, 95% CI: [57%, 64%]). Instructors performed significantly better than students (Welch's two-sample *t*-test: $t[283] = 3.30$, $p = .001$).

Familiarity With subject matter Participants rated how much previous knowledge they had in the essay subject matter (i.e., cognitive psychology and the science of learning). Linear regression models with *F*-tests indicated that familiarity with the subject did not predict instructors' or students' accuracy, $Fs(1) < 0.49$, $ps > .486$. Psychology instructors did not perform any better than non-psychology instructors, $t(130) = 0.18$, $p = .860$.

Familiarity with ChatGPT Nearly all participants (94%) said they had heard of ChatGPT before taking the survey, and most instructors (62%) and about half of students (50%) said they had used ChatGPT before. For both groups, participants who used ChatGPT did not perform any better than those who never used it before, $Fs(1) < 0.77$, $ps > .383$. Instructors' and students' experience with ChatGPT (from 0 = not at all experienced to 100 = extremely experienced) also did not predict their performance, $Fs(1) < 0.77$, $ps > .383$.

Confidence and estimated score Before they began the Identification Test, both instructors and students expressed low confidence in their abilities to identify the chatbot ($M = 34.60$ on a scale from 0 = not at all confident to 100 = extremely confident). Their confidence was significantly below the midpoint of the scale (midpoint: 50), one-sample *t*-test: $t(282) = 11.46$, $p < .001$, 95% CI: [31.95, 37.24]. Confidence ratings that were done before the AI Identification test did not predict performance for either group, Pearson's $rs < .12$, $ps > .171$.

Right after they completed the Identification Test, participants guessed how many text pairs they got right. Both instructors and students significantly underestimated their performance by about 15%, 95% CI: [11%, 18%], $t(279) = -8.42$, $p < .001$. Instructors' estimated scores were positively correlated with their actual scores, Pearson's $r = .20$,

$t(135) = 2.42$, $p = .017$. Students' estimated scores were not related to their actual scores, $r = .03$, $p = .731$.

Trial-by-trial performance on AI identification test

Confidence Participants' confidence ratings on individual trials were counted as high if they fell above the midpoint (>50 on a scale from 0=not at all confident to 100=extremely confident). For these within-subjects trial-by-trial analyses, we used Generalized Linear Mixed Models (GLMMs, Hox 2010) with random intercepts for participants and likelihood ratio tests (difference score reported as *D*). Both instructors and students performed better on trials in which they expressed high confidence (instructors: 73%, students: 63%) compared to low confidence (instructors: 65%, students: 56%), Ds(1) > 4.59, ps < .032.

Student essay quality We used two measures to capture the quality of each student-written essay: its assigned grade from 0 to 10 based on the class rubric, and its writing quality score from 0 to 100. Assigned grade was weakly related to instructors' accuracy, but not to students' accuracy. The text pairs that instructors got right tended to include student essays that earned slightly lower grades ($M = 7.89$, $SD = 2.22$) compared to those they got wrong ($M = 8.17$, $SD = 2.16$), $D(1) = 3.86$, $p = .050$. There was no difference for students, $D(1) = 2.84$, $p = .092$. Writing quality score did not differ significantly between correct and incorrect trials for either group, $D(1) = 2.12$, $p = .146$.

AI attitude assessment

Concerns and hopes about ChatGPT Both instructors and students expressed intermediate levels of concern and optimism. Specifically, on a scale from 0 ("not at all") to 100 ("extremely"), participants expressed intermediate concern about ChatGPT having negative effects on education ($M_{\text{instructors}} = 59.82$, $M_{\text{students}} = 55.97$) and intermediate optimism about it having positive benefits ($M_{\text{instructors}} = 49.86$, $M_{\text{students}} = 54.08$). Attitudes did not differ between instructors and students, ts < 1.43, ps > .154. Participants estimated that just over half of college students (instructors: 57%, students: 54%) would use ChatGPT to write an essay for them and submit it. These estimates also did not differ by group, $t(278) = 0.90$, $p = .370$.

Evaluations of ChatGPT uses Participants evaluated five different uses of ChatGPT in educational settings on a scale from −10 ("really bad") to +10 ("really good"). Both instructors and students rated it very bad for someone to ask ChatGPT to write an essay for them and submit the direct output, but instructors rated it significantly more negatively (instructors: -8.95, students: -7.74), $t(280) = 3.59$, $p < .001$. Attitudes did not differ between groups for any of the other scenarios (Table 2), ts < 1.31, ps > .130.

Exploratory analysis of demographic factors

We also conducted exploratory analyses looking at ChatGPT use and attitudes among different demographic groups (gender, race, and native English speakers). We combined instructors and students because their responses to the Attitude Assessment did not differ. In these exploratory analyses, we found that participants who were not native English speakers were more likely to report using ChatGPT and to view it more positively. Specifically, 69% of non-native English speakers had used ChatGPT before, versus 48%

Table 2 *Ratings of Hypothetical ChatGPT Uses in Academic Contexts*

		Evaluative Rating (*SD*)		
Scenario	**Text**	**Instructor**	**Student**	***p***
Direct	A student uses Chat GPT to write an essay for them and submits the direct generated answer.	-8.95 (1.88)	-7.74 (3.60)	< .001
Modify	A student uses Chat GPT to write an essay for their class, then the student edits the output and submits the revised essay.	-3.95 (4.50)	-3.21 (4.95)	.190
Enhance	A student writes an essay, asks Chat GPT to improve the essay's vocabulary or structure, and then submits the exact revision.	-0.73 (4.83)	-0.33 (4.70)	.477
Format	A student writes an essay, then uses Chat GPT to change it into a specific format (e.g., MLA, APA) and submits the output.	1.80 (5.67)	1.86 (5.02)	.930
Practice	A student uses Chat GPT to generate practice problems while studying for class.	5.88 (4.30)	6.51 (4.43)	.225

Note Evaluative rating was measured on a sliding scale from −10 = "really bad" to 0 = "neutral" to +10 = "really good." Significance was tested using Welch's two-sample *t*-tests

of native English speakers, $D(1)=12.00$, $p<.001$. Regardless of native language, the more experience someone had with ChatGPT, the more optimism they reported, $F(1)=18.71$, $p<.001$, $r=.37$). Non-native speakers rated the scenario where a student writes an essay and asks ChatGPT to improve its vocabulary slightly positively (1.19) whereas native English speakers rated it slightly negatively (-1.43), $F(1)=11.00$, $p=.001$. Asian participants expressed higher optimism ($M=59.14$) than non-Asian participants ($M=47.29$), $F(1)=10.05$, $p=.002$. We found no other demographic differences.

Study 2: ChatGPT

Study 1 provided data on college instructors' and students' ability to recognize ChatGPT-generated writing and about their views of the technology. In Study 2, of primary interest was whether ChatGPT itself might perform better at identifying ChatGPT-generated writing. Indeed, the authors have heard discussions of this as a possible solution to recognize AI-generated writing. We addressed this question by repeatedly asking ChatGPT to act as a participant in the AI Identification Task. While doing so, we administered the rest of the assessment given to participants in Study 1. This included our AI Attitude Assessment, which allowed us to examine the extent to which ChatGPT produced attitude responses that were similar to those of the participants in Study 1.

Participants, materials, and procedures

There were no human participants for Study 2. We collected 40 survey responses from ChatGPT, each run in a separate session on the platform (https://chat.openai.com/) between 5/4/2023 and 5/15/2023.

Two research assistants were trained on how to run the survey in the ChatGPT online interface. All prompts from the Study 1 survey were used, with minor modifications to suit the chat format. For example, slider questions were explained in the prompt, so instead of "How confident are you about this answer?" the prompt was "How confident are you about this answer from 0 (not at all confident) to 100 (extremely confident)?". In pilot testing, we found that ChatGPT sometimes failed to answer the question (e.g., by not providing a number), so we prepared a second prompt for every question that the researcher used whenever the first prompt was not answered (e.g., "Please answer the above question with one number between 0 to 100."). If ChatGPT still failed on the second prompt, the researcher marked it as a non-response and moved on to the next question in the survey.

Data analysis Like Study 1, all analyses were done in R Statistical Software (R Core Team 2021). Key analyses first used linear regression models and *F*-tests to compare all three groups (instructors, students, ChatGPT). When these omnibus tests were significant, we followed up with post-hoc pairwise comparisons using Tukey's method.

Results

AI identification test

Overall accuracy ChatGPT generated correct responses on 63% of trials in the AI Identification Test, which was significantly above chance, binomial test $p < .001$, 95% CI: [57%, 69%]. Pairwise comparisons found that this performance by ChatGPT was not any different from that of instructors or students, $ts(322) < 1.50$, $ps > .292$.

Confidence and estimated performance Unlike the human participants, ChatGPT produced responses with very high confidence before the task generally ($m = 71.38$, $median = 70$) and during individual trials specifically ($m = 89.82$, $median = 95$). General confidence ratings before the test were significantly higher from ChatGPT than from the humans (instructors: 34.35, students: 34.83), $ts(320) > 9.47$, $ps < .001$. But, as with the human participants, this confidence did not predict performance on the subsequent Identification task, $F(1) = 0.94$, $p = .339$. And like the human participants, ChatGPT's reported confidence on individual trials did predict performance: ChatGPT produced higher confidence ratings on correct trials ($m = 91.38$) than incorrect trials ($m = 87.33$), $D(1) = 8.74$, $p = .003$.

ChatGPT also produced responses indicating high confidence after the task, typically estimating that it got all six text pairs right ($M = 91\%$, $median = 100\%$). It overestimated performance by about 28%, and a paired *t*-test confirmed that ChatGPT's estimated performance was significantly higher than its actual performance, $t(36) = 9.66$, $p < .001$. As inflated as it was, estimated performance still had a small positive correlation with actual performance, Pearson's $r = .35$, $t(35) = 2.21$, $p = .034$.

Essay quality The quality of the student essays as indexed by their grade and writing quality score did not significantly predict performance, $Ds < 1.97$, $ps > .161$.

AI attitude Assessment

Concerns and hopes ChatGPT usually failed to answer the question, "How concerned are you about ChatGPT having negative effects on education?" from 0 (not at all concerned) to 100 (extremely concerned). Across the 40% of cases where ChatGPT successfully produced an answer, the average concern rating was 64.38, which did not differ significantly from instructors' or students' responses, $F(2, 294) = 1.20$, $p = .304$. ChatGPT produced answers much more often for the question, "How optimistic are you about ChatGPT having positive benefits for education?", answering 88% of the time. The average optimism rating produced by ChatGPT was 73.24, which was significantly higher than that of instructors (49.86) and students (54.08), $ts > 4.33$, $ps < .001$. ChatGPT only answered 55% of the time for the question about how many students would use ChatGPT to write an essay for them and submit it, typically generating explanations about its inability to predict human behavior and the fact that it does not condone cheating when it did

not give an estimate. When it did provide an estimate ($m = 10\%$), it was vastly lower than that of instructors (57%) and students (54%), $ts > 7.84$, $ps < .001$.

Evaluation of ChatGPT uses ChatGPT produced ratings of the ChatGPT use scenarios that on average were rank-ordered the same as the human ratings, with direct copying rated the most negatively and generating practice problems rated the most positively (see Fig. 2).

Compared to humans' ratings, ratings produced by ChatGPT were significantly more positive in most scenarios, $ts > 3.09$, $ps < .006$, with two exceptions. There was no significant difference between groups in the "format" scenario (using ChatGPT to format an essay in another style such as APA), $F(2,318) = 2.46$, $p = .087$. And for the "direct" scenario, ChatGPT tended to rate direct copying more negatively than students ($t[319] = 4.08$, $p < .001$) but not instructors ($t[319] = 1.57$, $p = .261$), perhaps because ratings from ChatGPT and instructors were already so close to the most negative possible rating.

Discussion

In 1950, Alan Turing said he hoped that one day machines would be able to compete with people in all intellectual fields (Turing 1950; see Köbis and Mossink 2021). Today, by many measures, the large-language model, ChatGPT, appears to be getting close to achieving this end. In doing so, it is raising questions about the impact this AI and its successors will have on individuals and the institutions that shape the societies in which we live. One important set of questions revolves around its use in higher education, which is the focus of the present research.

Empirical contributions

Detecting AI-generated text

Our central research question focused on whether instructors can identify ChatGPT-generated writing, since an inability to do so could threaten the ability of institutions of higher learning to promote learning and assess competence. To address this question, we developed an AI Identification Test in which the goal was to try to distinguish between psychology essays written by college students on exams versus essays generated

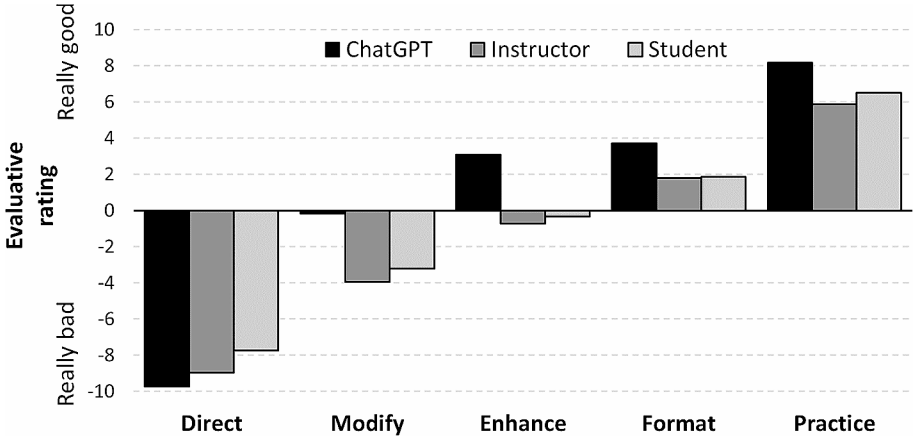

Fig. 2 Average ratings of ChatGPT uses, from − 10 = really bad to + 10 = really good. Human responses included for comparison (instructors in dark gray and students in light gray bars)

by ChatGPT in response to the same prompts. We found that although college instructors performed substantially better than chance, they still found the assessment to be challenging, scoring an average of only 70%. This relatively poor performance suggests that college instructors have substantial difficulty detecting ChatGPT-generated writing. Interestingly, this performance by the college instructors was the same average performance as Waltzer et al. (2023a) observed among high school instructors (70%) on a similar test involving English literature essays, suggesting the results are generalizable across the student populations and essay types. We also gave the assessment to college students (Study 1) and to ChatGPT (Study 2) for comparison. On average, students (60%) and ChatGPT (63%) performed even worse than instructors, although the difference only reached statistical significance when comparing students and instructors.

We found that instructors and students who went into the study believing they would be very good at distinguishing between essays written by college students versus essays generated by ChatGPT were in fact no better at doing so than participants who lacked such confidence. However, we did find that item-level confidence did predict performance: when participants rated their confidence after each specific pair (i.e., "How confident are you about this answer?"), they did perform significantly better on items they reported higher confidence on. These same patterns were observed when analyzing the confidence ratings from ChatGPT, though ChatGPT produced much higher confidence ratings than instructors or students, reporting overconfidence while instructors and students reported underconfidence.

Attitudes toward AI in education

Instructors and students both thought it was very bad for students to turn in an assignment generated by ChatGPT as their own, and these ratings were especially negative for instructors. Overall, instructors and students looked similar to one another in their evaluations of other uses of ChatGPT in education. For example, both rated submitting an edited version of a ChatGPT-generated essay in a class as bad, but less bad than submitting an unedited version. Interestingly, the rank orderings in evaluations of ChatGPT uses were the same when the responses were generated by ChatGPT as when they were generated by instructors or students. However, ChatGPT produced more favorable ratings of several uses compared to instructors and students (e.g., using the AI tool to enhance the vocabulary in an essay). Overall, both instructors and students reported being about as optimistic as they were concerned about AI in education. Interestingly, ChatGPT produced responses indicative of much more optimism than both human groups of participants.

Many instructors commented on the challenges ChatGPT poses for educators. One noted that "… ChatGPT makes it harder for us to rely on homework assignments to help students to learn. It will also likely be much harder to rely on grading to signal how likely it is for a student to be good at a skill or how creative they are." Some suggested possible solutions such as coupling writing with oral exams. Others suggested that they would appreciate guidance. For example, one said, "I have told students not to use it, but I feel like I should not be like that. I think some of my reluctance to allow usage comes from not having good guidelines."

And like the instructors, some students also suggested that they want guidance, such as knowing whether using ChatGPT to convert a document to MLA format would count

as a violation of academic integrity. They also highlighted many of the same problems as instructors and noted beneficial ways students are finding to use it. One student noted that, "I think ChatGPT definitely has the potential to be abused in an educational setting, but I think at its core it can be a very useful tool for students. For example, I've heard of one student giving ChatGPT a rubric for an assignment and asking it to grade their own essay based on the rubric in order to improve their writing on their own."

Theoretical contributions and practical implications

Our findings underscore the fact that AI chatbots have the potential to produce confident-sounding responses that are misleading (Chen et al. 2023; Goodwins 2022; Salvi et al. 2024). Interestingly, the underconfidence reported by instructors and students stands in contrast to some findings that people often expressed overconfidence in their abilities to detect AI (e.g., deepfake videos, Köbis et al. 2021). Although general confidence before the task did not predict performance, specific confidence on each item of the task *did* predict performance. Taken together, our findings are consistent with other work suggesting confidence effects are context-dependent and can differ depending on whether they are assessed at the item level or more generally (Gigerenzer et al. 1991).

The fact that college instructors have substantial difficulty differentiating between ChatGPT-generated writing and the writing of college students provides evidence that ChatGPT poses a significant threat to academic integrity. Ignoring this threat is also likely to undermine central aspects of the mission of higher education in ways that undermine the value of assessments and disincentivize the kinds of cognitive engagement that promote deep learning (Chi and Wylie 2014). We are skeptical of answers that point to the use of AI detection tools to address this issue given that they will always be imperfect and false accusations have potential to cause serious harm (Dalalah and Dalalah 2023; Fowler 2023; Svrluga, 2023). Rather, we think that the solution will have to involve developing and disseminating best practices regarding creating assessments and incentivizing cognitive engagement in ways that help students learn to use AI as problem-solving tools.

Limitations and future directions

Why instructors perform better than students at detecting AI-generated text is unclear. Although we did not find any effect of content-relevant expertise, it still may be the case that experience with evaluating student writing matters, and instructors presumably have more such experience. For example, one non-psychology instructor who got 100% of the pairs correct said, "Experience with grading lower division undergraduate papers indicates that students do not always fully answer the prompt, if the example text did not appear to meet all of the requirements of the prompt or did not provide sufficient information, I tended to assume an actual student wrote it." To address this possibility, it will be important to compare adults who do have teaching experience with those who do not.

It is somewhat surprising that experience with ChatGPT did not affect the performance of instructors or students on the AI Identification Test. One contributing factor may be that people pick up on some false heuristics from reading the text it generates (see Jakesch et al. 2023). It is possible that giving people practice at distinguishing the different forms of writing with feedback could lead to better performance.

Why confidence was predictive of accuracy at the item level is still not clear. One possibility is that there are some specific and valid cues many people were using. One likely cue is grammar. We revised grammar errors in student essays that were picked up by a standard spell checker in which the corrections were obvious. However, we left ungrammatical writing that didn't have obvious corrections (e.g., "That is being said, to be able to understand the concepts and materials being learned, and be able to produce comprehension."). Many instructors noted that they used grammatical errors as cues that writing was generated by students. As one instructor remarked, "Undergraduates often have slight errors in grammar and tense or plurality agreement, and I have heard the chat bot works very well as an editor." Similarly, another noted, "I looked for more complete, grammatical sentences. In my experience, Chat-GPT doesn't use fragment sentences and is grammatically correct. Students are more likely to use incomplete sentences or have grammatical errors." This raises methodological questions about what is the best comparison between AI and human writing. For example, it is unclear which grammatical mistakes should be corrected in student writing. Also of interest will be to examine the detectability of writing that is generated by AI and later edited by students, since many students will undoubtedly use AI in this way to complete their course assignments.

We also found that student-written essays that earned higher grades (based on the scoring rubric for their class exam) were harder for instructors to differentiate from ChatGPT writing. This does not appear to be a simple effect of writing quality given that a separate measure of writing quality that did not account for content accuracy was not predictive. According to the class instructor, the higher-scoring essays tended to include more specific details, and this might have been what made them less distinguishable. Relatedly, it may be that the higher-scoring essays were harder to distinguish because they appeared to be generated by more competent-sounding writers, and it was clear from instructor comments that they generally viewed ChatGPT as highly competent.

Conclusion

The results of the present research validate concerns that have been raised about college instructors having difficulty distinguishing writing generated by ChatGPT from the writing of their students, and document that this is also true when students try to detect writing generated by ChatGPT. The results indicate that this issue is particularly pronounced when instructors evaluate high-scoring student essays. The results also indicate that ChatGPT itself performs no better than instructors at detecting ChatGPT-generated writing even though ChatGPT-reported confidence is much higher. These findings highlight the importance of examining current teaching and assessment practices and the potential challenges AI chatbots pose for academic integrity and ethics in education (Cotton et al. 2023; Eke 2023; Susnjak 2022). Further, the results show that both instructors and students have a mixture of apprehension and optimism about the use of AI in education, and that many are looking for guidance about how to ethically use it in ways that promote learning. Taken together, our findings underscore some of the challenges that need to be carefully navigated in order to minimize the risks and maximize the benefits of AI in education.

Abbreviations

AI	Artificial Intelligence
CI	Confidence Interval
GLMM	Generalized Linear Mixed Model

GPT Generative Pre-trained Transformer
SD Standard Deviation

Acknowledgements
We thank Daniel Chen and Riley L. Cox for assistance with study design, stimulus preparation, and pilot testing. We also thank Emma C. Miller for grading the essays and Brian J. Compton for comments on the manuscript.

Author contributions
All authors collaborated in the conceptualization and design of the research. C. Pilegard facilitated recruitment and coding for real class assignments used in the study. T. Waltzer led data collection and analysis. G. Heyman and T. Waltzer wrote and revised the manuscript.

Funding
This work was partly supported by a National Science Foundation Postdoctoral Fellowship for T. Waltzer (NSF SPRF-FR# 2104610).

Data availability
Supplementary materials, including data, analysis, and survey items, are available on the Open Science Framework: https://osf.io/2c54a/.

Received: 16 February 2024 / Accepted: 11 June 2024
Published online: 08 July 2024

References

Al Darayseh A (2023) Acceptance of artificial intelligence in teaching science: Science teachers' perspective. Computers Education: Artif Intell 4:100132. https://doi.org/10.1016/j.caeai.2023.100132
Bertram Gallant T (2011) Creating the ethical academy. Routledge, New York
Biswas SS (2023) Potential use of Chat GPT in global warming. Ann Biomed Eng 51:1126–1127. https://doi.org/10.1007/s10439-023-03171-8
Borenstein J, Howard A (2021) Emerging challenges in AI and the need for AI ethics education. AI Ethics 1:61–65. https://doi.org/10.1007/s43681-020-00002-7
Bretag T (ed) (2016) Handbook of academic integrity. Springer
Bretag T, Harper R, Burton M, Ellis C, Newton P, Rozenberg P, van Haeringen K (2019) Contract cheating: a survey of Australian university students. Stud High Educ 44(11):1837–1856. https://doi.org/10.1080/03075079.2018.1462788
Brown TB, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler D, Wu J, Winter C, Amodei D (2020) Language models are few-shot learners. Adv Neural Inf Process Syst 33. https://doi.org/10.48550/arxiv.2005.14165
Chen Y, Andiappan M, Jenkin T, Ovchinnikov A (2023) A manager and an AI walk into a bar: does ChatGPT make biased decisions like we do? SSRN 4380365. https://doi.org/10.2139/ssrn.4380365
Chi MTH, Wylie R (2014) The ICAP framework: linking cognitive engagement to active learning outcomes. Educational Psychol 49(4):219–243. https://doi.org/10.1080/00461520.2014.965823
Chocarro R, Cortiñas M, Marcos-Matás G (2023) Teachers' attitudes towards chatbots in education: a technology acceptance model approach considering the effect of social language, bot proactiveness, and users' characteristics. Educational Stud 49(2):295–313. https://doi.org/10.1080/03055698.2020.1850426
Cizek GJ (1999) Cheating on tests: how to do it, detect it, and prevent it. Routledge
R Core Team (2021) *R: A language and environment for statistical computing* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
Cotton DRE, Cotton PA, Shipway JR (2023) Chatting and cheating: ensuring academic integrity in the era of ChatGPT. Innovations Educ Teach Int. https://doi.org/10.1080/14703297.2023.2190148
Curtis GJ, Clare J (2017) How prevalent is contract cheating and to what extent are students repeat offenders? J Acad Ethics 15:115–124. https://doi.org/10.1007/s10805-017-9278-x
Dalalah D, Dalalah OMA (2023) 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(2):100822. https://doi.org/10.1016/j.ijme.2023.100822
Devlin J, Chang M-W, Lee K, Toutanova K (2018) BERT: pre-training of deep bidirectional transformers for language understanding. ArXiv. https://doi.org/10.48550/arxiv.1810.04805
Dwivedi YK, Kshetri N, Hughes L, Slade EL, Jeyaraj A, Kar AK, Baabdullah AM, Koohang A, Raghavan V, Ahuja M, Albanna H, Albashrawi MA, Al-Busaidi AS, Balakrishnan J, Barlette Y, Basu S, Bose I, Brooks L, Buhalis D, Wright R (2023) So what if ChatGPT wrote it? Multidisciplinary perspectives on opportunities, challenges, and implications of generative conversational AI for research, practice, and policy. Int J Inf Manag 71:102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642
Eke DO (2023) ChatGPT and the rise of generative AI: threat to academic integrity? J Responsible Technol 13:100060. https://doi.org/10.1016/j.jrt.2023.100060
Erickson S, Heit E (2015) Metacognition and confidence: comparing math to other academic subjects. Front Psychol 6:742. https://doi.org/10.3389/fpsyg.2015.00742
Fischer I, Budescu DV (2005) When do those who know more also know more about how much they know? The development of confidence and performance in categorical decision tasks. Organ Behav Hum Decis Process 98:39–53. https://doi.org/10.1016/j.obhdp.2005.04.003
Fleming SM, Weil RS, Nagy Z, Dolan RJ, Rees G (2010) Relating introspective accuracy to individual differences in brain structure. Science 329:1541–1543. https://doi.org/10.1126/science.1191883
Fowler GA (2023), April 14 We tested a new ChatGPT-detector for teachers. It flagged an innocent student. *The Washington Post*. https://www.washingtonpost.com/technology/2023/04/01/chatgpt-cheating-detection-turnitin/

Gigerenzer G (1991) From tools to theories: a heuristic of discovery in cognitive psychology. Psychol Rev 98:254. https://doi.org/10.1037/0033-295X.98.2.254
Gigerenzer G, Hoffrage U, Kleinbölting H (1991) Probabilistic mental models: a brunswikian theory of confidence. Psychol Rev 98(4):506–528. https://doi.org/10.1037/0033-295X.98.4.506
Gilson A, Safranek C, Huang T, Socrates V, Chi L, Taylor RA, Chartash D (2022) How well does ChatGPT do when taking the medical licensing exams? The implications of large language models for medical education and knowledge assessment. MedRxiv. https://doi.org/10.1101/2022.12.23.22283901
Goodwins T (2022), December 12 ChatGPT has mastered the confidence trick, and that's a terrible look for AI. *The Register*. https://www.theregister.com/2022/12/12/chatgpt_has_mastered_the_confidence/
Gunser VE, Gottschling S, Brucker B, Richter S, Gerjets P (2021) Can users distinguish narrative texts written by an artificial intelligence writing tool from purely human text? In C. Stephanidis, M. Antona, & S. Ntoa (Eds.), *HCI International 2021, Communications in Computer and Information Science*, (Vol. 1419, pp. 520–527). Springer. https://doi.org/10.1007/978-3-030-78635-9_67
Hartshorne H, May MA (1928) Studies in the nature of character: vol. I. studies in deceit. Macmillan, New York
Hox J (2010) Multilevel analysis: techniques and applications, 2nd edn. Routledge, New York, NY
Jakesch M, Hancock JT, Naaman M (2023) Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences, 120*(11), e2208839120. https://doi.org/10.1073/pnas.2208839120
Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, Wang Y, Dong Q, Shen H, Wang Y (2017) Artificial intelligence in healthcare: past, present and future. Stroke Vascular Neurol 2(4):230–243. https://doi.org/10.1136/svn-2017-000101
Joo YJ, Park S, Lim E (2018) Factors influencing preservice teachers' intention to use technology: TPACK, teacher self-efficacy, and technology acceptance model. J Educational Technol Soc 21(3):48–59. https://www.jstor.org/stable/26458506
Kasneci E, Seßler K, Küchemann S, Bannert M, Dementieva D, Fischer F, Kasneci G (2023) ChatGPT for good? On opportunities and challenges of large language models for education. Learn Individual Differences 103:102274. https://doi.org/10.1016/j.lindif.2023.102274
Katz DM, Bommarito MJ, Gao S, Arredondo P (2023) GPT-4 passes the bar exam. SSRN Electron J. https://doi.org/10.2139/ssrn.4389233
Köbis N, Mossink LD (2021) Artificial intelligence versus Maya Angelou: experimental evidence that people cannot differentiate AI-generated from human-written poetry. Comput Hum Behav 114:106553. https://doi.org/10.1016/j.chb.2020.106553
Köbis NC, Doležalová B, Soraperra I (2021) Fooled twice: people cannot detect deepfakes but think they can. iScience 24(11):103364. https://doi.org/10.1016/j.isci.2021.103364
Lo CK (2023) What is the impact of ChatGPT on education? A rapid review of the literature. Educ Sci 13(4):410. https://doi.org/10.3390/educsci13040410
McCabe DL, Butterfield KD, Treviño LK (2012) Cheating in college: why students do it and what educators can do about it. Johns Hopkins, Baltimore, MD
Mitchell A (2022) December 26). Professor catches student cheating with ChatGPT: 'I feel abject terror'. New York Post. https://nypost.com/2022/12/26/students-using-chatgpt-to-cheat-professor-warns
Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I (2019) Language models are unsupervised multitask learners. OpenAI https://openai.com/research/better-language-models
Rettinger DA, Bertram Gallant T (eds) (2022) Cheating academic integrity: lessons from 30 years of research. Jossey Bass
Rosenzweig-Ziff D (2023) New York City blocks use of the ChatGPT bot in its schools. Wash Post https://www.washingtonpost.com/education/2023/01/05/nyc-schools-ban-chatgpt/
Salvi F, Ribeiro MH, Gallotti R, West R (2024) On the conversational persuasiveness of large language models: a randomized controlled trial. ArXiv. https://doi.org/10.48550/arXiv.2403.14380
Shynkaruk JM, Thompson VA (2006) Confidence and accuracy in deductive reasoning. Mem Cognit 34(3):619–632. https://doi.org/10.3758/BF03193584
Stokel-Walker C (2022) AI bot ChatGPT writes smart essays — should professors worry? Nature. https://doi.org/10.1038/d41586-022-04397-7
Susnjak T (2022) ChatGPT: The end of online exam integrity? *ArXiv*. https://arxiv.org/abs/2212.09292
Svrluga S (2023) Princeton student builds app to detect essays written by a popular AI bot. Wash Post https://www.washingtonpost.com/education/2023/01/12/gptzero-chatgpt-detector-ai/
Terwiesch C (2023) Would Chat GPT3 get a Wharton MBA? A prediction based on its performance in the Operations Management course. *Mack Institute for Innovation Management at the Wharton School*, University of Pennsylvania. https://mackinstitute.wharton.upenn.edu/wp-content/uploads/2023/01/Christian-Terwiesch-Chat-GTP-1.24.pdf
Tlili A, Shehata B, Adarkwah MA, Bozkurt A, Hickey DT, Huang R, Agyemang B (2023) What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. Smart Learn Environ 10:15. https://doi.org/10.1186/s40561-023-00237-x
Turing AM (1950) Computing machinery and intelligence. Mind - Q Rev Psychol Philos 236:433–460
UCSD Academic Integrity Office (2023) GenAI, cheating and reporting to the AI office [Announcement]. https://adminrecords.ucsd.edu/Notices/2023/2023-5-17-1.html
Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst 30. https://doi.org/10.48550/arxiv.1706.03762
Waltzer T, Dahl A (2023) Why do students cheat? Perceptions, evaluations, and motivations. Ethics Behav 33(2):130–150. https://doi.org/10.1080/10508422.2022.2026775
Waltzer T, Cox RL, Heyman GD (2023a) Testing the ability of teachers and students to differentiate between essays generated by ChatGPT and high school students. Hum Behav Emerg Technol 2023:1923981. https://doi.org/10.1155/2023/1923981
Waltzer T, DeBernardi FC, Dahl A (2023b) Student and teacher views on cheating in high school: perceptions, evaluations, and decisions. J Res Adolescence 33(1):108–126. https://doi.org/10.1111/jora.12784
Weidinger L, Mellor J, Rauh M, Griffin C, Uesato J, Huang PS, Gabriel I (2021) Ethical and social risks of harm from language models. ArXiv. https://doi.org/10.48550/arxiv.2112.04359
Wixted JT, Wells GL (2017) The relationship between eyewitness confidence and identification accuracy: a new synthesis. Psychol Sci Public Interest 18(1):10–65. https://doi.org/10.1177/1529100616686966

Yeadon W, Inyang OO, Mizouri A, Peach A, Testrow C (2023) The death of the short-form physics essay in the coming AI revolution. Phys Educ 58:035027. https://doi.org/10.1088/1361-6552/acc5cf
Zhuo TY, Huang Y, Chen C, Xing Z (2023) Red teaming ChatGPT via jailbreaking: bias, robustness, reliability and toxicity. ArXiv. https://doi.org/10.48550/arxiv.2301.12867

Publisher's Note