International Journal of Artificial Intelligence in Education
https://doi.org/10.1007/s40593-024-00427-9

ARTICLE

ChatGPT in Undergraduate Education: Performance of GPT-3.5 and Identification of AI-Generated Text in Introductory Neuroscience

Natalie V. Covington[1,2] · **Olivia Vruwink**[1]

Accepted: 13 August 2024

Abstract

ChatGPT and other large language models (LLMs) have the potential to significantly disrupt common educational practices and assessments, given their capability to quickly generate human-like text in response to user prompts. LLMs GPT-3.5 and GPT-4 have been tested against many standardized and high-stakes assessment materials (e.g. SAT, Uniform Bar Exam, GRE), demonstrating impressive but variable performance. Fewer studies have examined the performance of ChatGPT on course-level educational materials in ecologically-valid grading contexts. Here, we examine the performance of GPT-3.5 on undergraduate course materials and assess the ability of teaching assistants to identify AI-generated responses interleaved with student work. GPT-3.5 was prompted to respond to questions drawn from undergraduate neuroscience assessments. These AI-generated responses were interleaved with student-authored responses and graded blindly using existing course rubrics. In addition, a subset of responses were rated for their humanlikeness by teaching assistants who were blind to author status (AI vs. student). In general, GPT-3.5 performed within one standard deviation of the class average, but there were cases in which ChatGPT-generated responses substantially outperformed or underperformed relative to student responses. Teaching assistants who were blind to author status were able to identify ChatGPT-generated responses with better than chance accuracy, and those with personal experience using ChatGPT were significantly more accurate than those without ChatGPT experience. Despite high levels of identification accuracy, none of the teaching assistant raters endorsed sufficient confidence in their identifications to support reporting the response as an instance of academic dishonesty in a real-world classroom setting.

Keywords Artificial intelligence · Large language models · Undergraduate education · Neuroscience

Extended author information available on the last page of the article

Published online: 09 September 2024

Introduction

ChatGPT was publicly released by OpenAI in November of 2022 and quickly surpassed existing records for most-downloaded application. The initial application, based on large language model (LLM) GPT-3.5, provided the general public with a glimpse at the remarkable advances in natural language processing that have occurred over the past five years. In contrast to older models, ChatGPT is capable of providing responses to a wide range of task-specific prompts with well-structured, humanlike prose without additional task-specific fine-tuning. These advances in the humanlikeness of ChatGPT's responses relative to previous models have been attributed in part to the number of parameters in the model (175 billion) and the massive data set on which it was trained (including text data from vast swaths of internet archives) (Brown et al., 2020).

Concerns about ChatGPT's impacts on learning and higher education emerged almost concurrently with its release. Early testing of ChatGPT demonstrated that GPT-3.5 was capable of strong performance on a variety of high-stakes and standardized assessments, including the SAT and several AP exams (OpenAI, 2023). A newer and more powerful ChatGPT model (GPT-4) demonstrated strong performance on the Uniform Bar and the GRE and outperformed GPT-3.5 on most of the high-stakes exams on which it was tested (OpenAI, 2023). Given this strong performance, educators have raised concerns about the potential for the use of ChatGPT and other AI-enhanced tools to facilitate academic misconduct (Perkins, 2023) and to reduce the rigor of common undergraduate assignments and assessments (Malik et al., 2023; Rudolph et al., 2023). Essays written by ChatGPT bypass detection by common plagiarism detectors (Khalil & Er, 2023) and are of high enough quality to pass graduate-level essay exams (Choi et al., 2023).

In light of these concerns, we sought to examine the performance of ChatGPT on undergraduate neuroscience course materials, using methods that reflect real-world contexts. Given ChatGPT's strong performance on high-stakes testing, and its passing, but middle-of-the-road, performance on graduate-level course material (Choi et al., 2023), we aimed to investigate its capability in course contexts that are more foundational. In addition, we examined the ability of teaching assistants to identify machine-generated text in ecologically-valid contexts. Our study adds to the emerging literature evaluating the performance of LLMs on undergraduate assessments and the capability of human readers to detect machine-generated text in real-world contexts. Understanding how ChatGPT performs on common undergraduate assessments and how its use may (or may not) be detected in contexts that approximate real-world courses is particularly critical given the major shifts in assessment practices initiated by the COVID-19 pandemic, in which a majority of instructors moved assessment online (Chan, 2022). Importantly, emerging evidence suggests that many faculty members have opted to retain these new assessment formats even after the return to in-person learning (Kerrigan et al., 2022).

Background

Performance of ChatGPT in Academic Contexts

Determining how well ChatGPT performs in academic contexts is important for understanding its likelihood of adoption by undergraduate students and its potential for disruption of common assignment and assessment methods in higher education. Here, we review recent work that tests the capability of ChatGPT across varied educational assessments, assignments, and levels.

Performance on High Stakes Exams OpenAI, the AI research organization behind ChatGPT, has used exams originally designed for human test takers as benchmarks against which to assess ChatGPT's performance and demonstrate its potential (OpenAI, 2023). Exams used for this purpose by OpenAI have primarily been comprised of "high stakes" or standardized academic and professional exams, ranging from Advanced Placement exams (e.g. AP US Government, AP Microeconomics), higher education entrance exams (e.g. the LSAT, SAT, GRE), and professional exams (e.g. Medical Knowledge Self-Assessment Program; Sommelier exams; OpenAI, 2023). For each reported exam, responses by GPT-3.5 and GPT-4 were scored using test-specific scoring methodologies and compared to estimated human score distributions (i.e. GPT-authored test performance reported as a percentile within the human score distribution). ChatGPT's performance on these high stakes exams is variable (ranging from below the 5th percentile to above the 90th percentile), with GPT-4 typically outperforming GPT-3.5 (OpenAI, 2023). These reports demonstrate the general-purpose power of ChatGPT and highlight GPT-3.5/GPT-4's advances relative to previous LLMs. The high stakes exams against which ChatGPT has been tested include a mix of multiple choice and free response questions, but data reported by OpenAI are limited to overall exam scores and percentile benchmarks and so the utility of these reports for informing decision-making in undergraduate course contexts is limited.

Performance on Higher Education Course Material Beyond high stakes assessments, performance of ChatGPT in "everyday" higher education contexts has been evaluated via prompting GPT-3.5 or GPT-4 using course materials ranging from undergraduate multiple choice quizzes and essay assignments to graduate-level final exams. Choi and colleagues provided an early demonstration of both the potential disruptive power of ChatGPT and also of GPT-3.5's limitations by assessing its performance on law school examinations (Choi et al., 2023). In this study, GPT-3.5 was prompted with essay and multiple-choice format questions that comprised the final exams for four courses within a JD law program. ChatGPT-authored responses were generated by a study team member and shuffled with actual student responses, and then blindly graded by other members of the study team. Across the four exams, ChatGPT-authored responses earned overall passing grades, with a grand average across the exams that was equivalent to a C + (Choi et al., 2023). The authors note that this level of performance would earn progress towards the JD degree, but would

also result in the ChatGPT "student" being placed on academic probation. Compared to human students in the course, ChatGPT-authored exams were consistently scored at the bottom of the class distribution; however, performance on individual questions was variable, with some ChatGPT-authored essay responses earning scores that exceeded student averages. In evaluating what factors led to higher or lower performance at the individual question level, Choi and colleagues note that ChatGPT struggled on essay questions that required reference to specific legal cases or to material specifically covered in class sessions (Choi et al., 2023). These results parallel findings that, when asked to cite specific sources, ChatGPT often hallucinates well-structured, but completely fabricated, academic citations (Buchanan et al., 2024; Walters & Wilder, 2023). On multiple choice questions, ChatGPT performed best when questions drew on universal legal principles that are applicable to most jurisdictions and struggled when questions required contextualization based on jurisdiction (Choi et al., 2023). In sum, while results of the Choi and colleagues study tempered some concerns related to the disruptive potential of ChatGPT in higher education, the level and complexity of exam questions limit the applicability of their findings to more foundational course contexts in undergraduate education.

At the undergraduate level, emerging work across fields demonstrates ChatGPT's variable performance on course assignments and assessments. Kortemeyer (2023) examined ChatGPT's performance on physics homework, clicker questions, programming exercises, and exams. Across these assignment and assessment formats, ChatGPT responses achieved between 47%-90% accuracy. Similarly, Clark and colleagues (2023) investigated ChatGPT's performance on a set of five chemistry problems. ChatGPT's performance was variable, averaging 46.5% accuracy, with a wide 0% to 100% range. In comparison, students (n = 182) in the course achieved an average of 16.3% accuracy prior to instruction and 51.5% after instruction. Across these two studies, ChatGPT's performance was compared either to a previous grading framework or scored by graders who were not blind to the student vs AI status of the responses, as student responses were written and ChatGPT's were in typed text.

Detectability of ChatGPT Output

University systems and individual instructors have responded to widespread access to LLMs like ChatGPT with a variety of policies, ranging from complete prohibition, to some allowable use with attribution, to unconditional adoption (Villasenor, 2023). Whether and how to detect the use of ChatGPT and other AI tools by undergraduate students on assignments and assessments is an ongoing debate in higher education, by both students (Holland & Ciachir, 2024) and faculty (Mamo et al., 2024); however, investigations involving plagiarism detection software suggest that existing plagiarism detection tools are not sufficient to accurately identify AI-generated responses (Walters, 2023; Weber-Wulff et al., 2023).

Human evaluators (e.g. instructors, teaching assistants) are another potential avenue for identifying use of ChatGPT in classroom contexts. Prior work shows variable success of human raters in accurately identifying AI-generated text. In a study by E. Clark et al. (2021), each of their participants (n = 780) rated five text

samples across a variety of writing genres, including stories, news articles, and recipes. Overall, they found that participants accurately distinguished between GPT-3 and human authored text only 49.9% of the time. Similarly, Jakesch et al. (2023) examined human detection of GPT-3 authored text in the hospitality, dating, and professional domains, finding that across all three contexts participants (n = 4,600) accurately identified AI-generated text 51.7% of the time. Together, these studies suggested significant challenges in human judgements of machine-generated text, with human raters performing no better than chance.

Limited recent work has evaluated the detectability of ChatGPT-authored responses in more ecologically-valid and classroom-based settings. In a study by Parker et al. (2024), researchers prompted GPT-4 to respond to various assessment prompts drawn from a range of undergraduate courses. GPT-4 responses were obtained using two different prompting conditions. When AI-authored responses were obtained by supplying ChatGPT with the assessment prompt and rubric, graduate teaching assistants accurately identified 40% of AI-authored responses. When AI-authored responses were further refined by employing prompt engineering (i.e. asking ChatGPT to make specific improvements to its responses), graduate teaching assistants accurately identified 60% of AI-authored responses. The authors suggested that the significantly increased length of the prompt-engineered responses may have served as a cue that contributed to the increased accuracy of identification (Parker et al., 2024).

The Current Study

In the current study, we explore the potential challenges posed by ChatGPT in undergraduate education by examining ChatGPT's performance on course materials drawn from an upper-level undergraduate neuroscience course compared to student-authored exam responses. In addition, we assess the ability of teaching assistant raters to detect AI-generated text and explore factors that may contribute to accurate humanlikeness judgments.

Our study expands on early findings related to the performance and detectability of ChatGPT output by more closely simulating real-world use of ChatGPT in the classroom. First, to develop AI-authored stimulus materials, assessment questions were posed directly to ChatGPT without significant prompt engineering, based on evidence that this is how students and non-experts with limited AI experience typically engage with the chat interface (Noy & Zhang, 2023; Woo et al., 2023; Zamfirescu-Pereira et al., 2023). Second, teaching assistant raters were blind to author status (AI vs. student) and also to the number of AI-authored responses in each set of assessment responses they were tasked with grading or rating. Including both student- and AI-authored stimuli in the grading set increases the ecological validity of our approach compared to previous studies which *only* included ChatGPT responses that were marked against existing grading schema (Fergus et al., 2023; Pursnani et al., 2023). Third, in our study, a small number of AI-authored responses were interleaved within a larger set of student-authored responses, to further improve ecological validity compared to previous AI-identification studies in which stimulus

items were balanced across author types (50% AI-authored, 50% human-authored, Clark et al., 2021; Jakesch et al., 2023). Fourth, in contrast to studies that examine *either* performance of ChatGPT on undergraduate assessments (e.g. Bordt & von Luxburg, 2023) *or* the ability of instructors to accurately identify AI-generated text (e.g. Alexander et al., 2023), our study simultaneously examines both ChatGPT's performance on undergraduate course materials and the ability of raters to accurately identify AI-generated text. Examining both LLM performance relative to undergraduate students and teaching assistant AI-detection capability allows us to examine whether response quality or other factors are related to detectability. Finally, we examine how experience within a specific course (e.g. prior teaching assistant experience) or with ChatGPT itself influence the accuracy of blinded raters' assessments of the humanlikeness of AI- and student-authored texts. Taken together, results of the current study will provide important insights into the potential impacts of ChatGPT on undergraduate education and will yield avenues for further investigation that may help to address the challenges this rising technology poses to traditional forms of student assessment.

Methods

Course Material Overview

Course materials were drawn from an upper-level undergraduate neuroscience course. The course is a requirement in the major, and covers foundational concepts in cellular, systems, and cognitive neuroscience. Course materials were drawn from two exams and four quizzes. For students in the course, exams were conducted in person, were closed-book, and included multiple choice and short-answer questions followed by three long-format essay questions. Quizzes were online, asynchronous, and open-book and consisted of multiple choice, true/false, fill-in-the-blank, and matching questions. On each quiz, students responded to twenty-five questions randomly selected by the course learning management system from a larger bank of questions developed by the first author.

Stimuli Development

Study procedures were reviewed by the University of Minnesota Institutional Review Board and deemed exempt from IRB review. All AI-generated materials were obtained using GPT-3.5 in February and March of 2023. All student-authored data was fully de-identified before inclusion in the study and is reported only in aggregate. Class averages for the four quizzes were obtained in aggregate from the course learning management system. To obtain student-authored exam responses, exams were first de-identified by the first author by removing existing cover sheets (the only portion of the exam containing identifiable information) and then randomly assigned a numeric "Author ID". Student-authored exams were further broken down

Table 1 Number of student-authored stimulus items by essay topic

Exam 1	Ion Channels and Pumps ($n = 56$)	Resting Potential ($n = 56$)	Action Potential ($n = 56$)
Exam 2	Synaptic Transmission ($n = 55$)	Semicircular Canals ($n = 55$)	Visual Field Deficits ($n = 55$)

into individual essay questions to be re-graded and compared against AI-authored responses (see Table 1).

For each essay question, two AI-authored exam responses were generated via the ChatGPT chat interface and then handwritten onto physical exam copies by two research assistants. ChatGPT's responses were transcribed verbatim, with the exception of the final "summary" paragraph that is common in ChatGPT's responses. In the "With History" condition, ChatGPT completed the entire exam sequentially in the same chat window (including multiple choice and short answer questions). In the "No History" condition, a new chat window was opened for each individual essay question, so that responses from ChatGPT were void of any contextual information.

Student- and AI-authored essay responses were sorted by essay prompt and interleaved in a random order to be graded and rated by teaching assistants. Graders and raters were kept blind to author status (Student versus AI) and were also blind to the number of AI-generated responses included in each essay set. There were six total essay prompts, drawn from two exams. Across all six essays, there were a total of fifty-eight unique student authors. The number of student-authored responses varied slightly, because of missing student data due to absence or alternative test forms (see Table 1).

Procedures for Grading and Rating

Exam Performance Grading

Essay responses (both student- and AI-authored) were provided to the second author (O. Vruwink) to grade using existing course rubrics. The second author had prior teaching assistant experience grading essays for the course but was not involved in grading the particular essay prompts included in this study. The second author was blind to author status (AI vs. student) and also was blind to the number of ChatGPT responses interleaved with student responses.

Exam Humanlikeness Ratings

A subset of ten responses to each essay (two AI-authored, eight student-authored) were rated for "humanlikeness" by six doctoral student teaching assistants using scales and methods previously employed in studies examining human identification of AI-generated text (Clark et al., 2021; Ippolito et al., 2020; Jakesch et al., 2023). Teaching assistants were blind to author status and to the number of ChatGPT responses interleaved with student responses. In addition, teaching assistants varied in terms of prior exposure to course materials. Two of the raters had been teaching

assistants in prior semesters of the course and were familiar with the essay prompts and course content but had not seen any of the current semester's exam data. The remaining four raters were doctoral students in the department with no prior experience in this particular course. Prior to rating the essay responses, teaching assistants were asked to respond to a set of questions characterizing their experience with ChatGPT and other similar AI-enhanced tools (e.g. Google Bard).

Teaching assistants were provided with a set of ten responses for each essay prompt to rate before moving on to the next essay prompt. The order in which prompts were rated was counterbalanced across teaching assistants. For each essay response in a set, teaching assistants were asked to provide a humanlikeness rating on a scale from one to four, where 1 was "Definitely Human Written"; 2 was "Possibly Human Written"; 3 was "Possibly Machine Generated"; and 4 was "Definitely Machine Generated" (Clark et al., 2021; Ippolito et al., 2020). Teaching assistants also provided a brief open response rationale for their ratings. After rating all six essay response sets, teaching assistants were asked to guess how many AI-generated responses were present in the dataset, and were asked to provide the Author IDs of the responses they thought were AI-generated. At the conclusion of all rating activities, raters were asked to indicate whether or not, in a real-world setting, they would feel confident enough in their assessment of which Author IDs were AI-generated to report those responses as potential instances of academic dishonesty.

Quiz Performance Grading

Student quizzes were graded immediately and automatically within the course learning management system (LMS) and class aggregate data was obtained for comparison with ChatGPT. To obtain grades for ChatGPT, each quiz was launched within the LMS. Quiz questions randomly selected by the LMS from the question bank were posed verbatim to ChatGPT along with their response options. ChatGPT-generated quiz responses were obtained separately from exam responses, with a new chat window opened for each quiz. In general, ChatGPT provided answers that closely matched a provided response option, which was then entered into the LMS for grading. In cases where ChatGPT attempted to provide multiple answers for a quiz question that only allowed a single response, the phrase "Choose one answer" was added to the prompt and posed to ChatGPT again.

Data Coding and Statistical Analysis

Grades assigned to ChatGPT on exams and quizzes were compared to the student-authored class average and standard deviation. As in Clark and colleagues (2021), humanlikeness ratings were scored for identification accuracy. For Student-authored responses, ratings of 1 (Definitely Human Written) or 2 (Possibly Human Written) were scored "correct", while ratings of 3 or 4 were scored "incorrect". Likewise, for ChatGPT-authored responses, ratings of 3 (Possibly Machine Generated) and 4 (Definitely Machine Generated) were scored "correct" while ratings of 1 or 2 were scored "incorrect". Analyses of item-level identification accuracy were conducted

using a logistic mixed effects model with by-item and by-rater random effects. This model predicts the log-odds of accurately identifying an author as "student" vs. "machine" across all items and raters, with a random effects structure that accounts for the nested structure of the data (i.e. 6 observations per item, 60 observations per rater). Coefficient estimates and p-values for fixed effects of interest are reported in the text, with full model specification and output in Appendix A. In addition, we examined the open-ended rationales provided by teaching assistants for their ratings of the ChatGPT responses, compared to Student-authored responses that were identified as Human Written with the highest and lowest degree of accuracy. Rationales were coded using the annotation scheme described in Clark and colleagues, with the addition of new codes where warranted by the data (2021).

Results

Performance of GPT-3.5

Exams

ChatGPT's performance, averaged across six essay prompts and across the No History and With History conditions, was 78%, compared to the overall student average of 74%. Grades by essay question and condition are presented in Table 2. In most instances, ChatGPT's performance was within one standard deviation of the student-authored class average, with two notable exceptions: 1) for "Action Potential", both of ChatGPT's responses earned a considerably (> 1.5 SD) higher grade than the student average, 2) for "Synaptic Transmission", the Chat GPT "No History" response earned a substantially lower grade than the student average. On Exam 1, seven students earned higher essay scores than either ChatGPT response (13% of students), while forty-four (79%) students earned lower essay scores than either ChatGPT response. On Exam 2, nineteen students earned higher essay scores than

Table 2 Grades by essay question and condition

		Student-Authored Average Grade (SD)	ChatGPT (No History)	ChatGPT (With History)
Exam 1	*Ion Channels and Pumps*	71% (12%)	68%	72%
	Resting Potential	76% (11%)	72%	76%
	Action Potential	69% (12%)	100%	88%
	Overall Grade	**72% (10%)**	**80%**	**79%**
Exam 2	*Synaptic Transmission*	82% (10%)	56%	80%
	Semicircular Canals	77% (11%)	76%	84%
	Visual Field Deficits	70% (17%)	84%	80%
	Overall Grade	**76% (10%)**	**72%**	**81%**

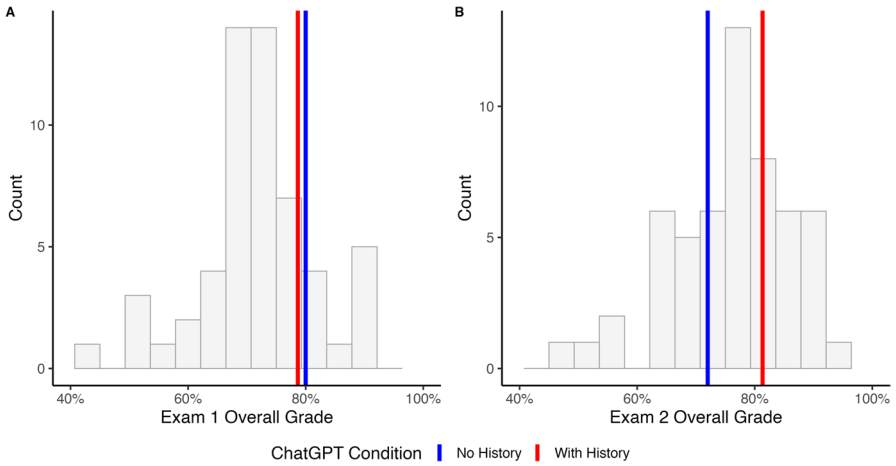

Fig. 1 Performance of ChatGPT relative to student grade distributions. Student grade distributions are depicted by gray histograms. ChatGPT "No History" performance is indicated with a blue vertical line. ChatGPT "With History" performance is indicated with a red vertical line

either ChatGPT response (35% of students), while fifteen (42%) students earned lower essay scores than either ChatGPT response. ChatGPT's performance on each exam relative to the student-authored distribution is presented in Fig. 1.

Quizzes

ChatGPT performed within a standard deviation of the class average in three out of four open-book course quizzes. Grades by quiz and author (student vs. ChatGPT) are presented in Table 3.

Identification of AI-Generated Text

Teaching Assistant Experience with AI

Half of the teaching assistants who completed humanlikeness ratings reported no prior experience using ChatGPT or any other AI-enhanced writing tool (e.g. Bard).

Table 3 Grades by quiz and author

	Student-Authored Average Grade (SD)	ChatGPT
Quiz 1	90% (6%)	94%
Quiz 2	92% (7%)	99%
Quiz 3	93% (6%)	86%
Quiz 4	87% (9%)	90%

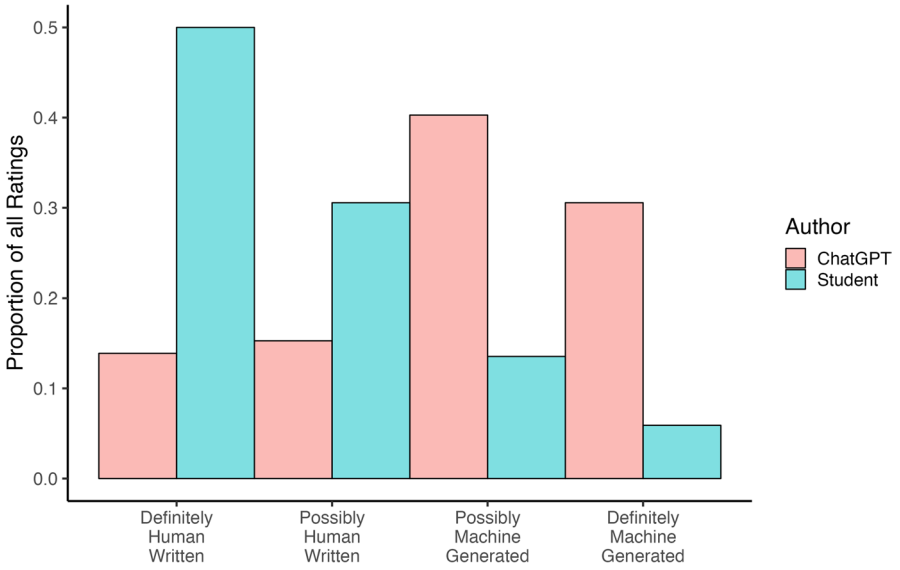

Fig. 2 Distribution of humanlikeness ratings by author

The remaining three teaching assistants reported having a ChatGPT account and using the service "once or twice a month". Teaching assistants with ChatGPT experience reported using the tool to draft emails, academic writing and tables, and as a tool to relieve writer's block and reword rough drafts.

Humanlikeness Ratings

Across all essays, ChatGPT-authored responses were given an average humanlikeness rating of 2.88, while Student-authored responses were given an average humanlikeness rating of 1.75 (where lower ratings indicate greater humanlikeness). The distribution of all teaching assistant ratings by Author (student vs. ChatGPT) is presented in Fig. 2.

Analysis of item-level identification accuracy was conducted using a logistic mixed effects model with by-item and by-rater random effects, and fixed effects of author type, AI experience, and course experience (see Appendix A for full model specification). Across all raters and essays, authors were identified with better than chance accuracy ($b = 0.51$, $p = 0.04$). Identification accuracy by author type was compared using Reverse Helmert contrasts, which compare each level of a categorical variable to the mean of previous levels. There was no difference in identification accuracy for the ChatGPT "With History" authored essays (identified with 83% accuracy) compared to Student-authored essays (identified with 81% accuracy; $b = 0.10$, $p = 0.68$). In contrast, the ChatGPT "No History" authored essays were identified with significantly poorer accuracy (58%; $b = -0.43$, $p = 0.003$). Figure 3 displays author identification accuracy based on rater experience. Rating accuracy was significantly higher for teaching assistants with AI experience ($b = 1.44$,

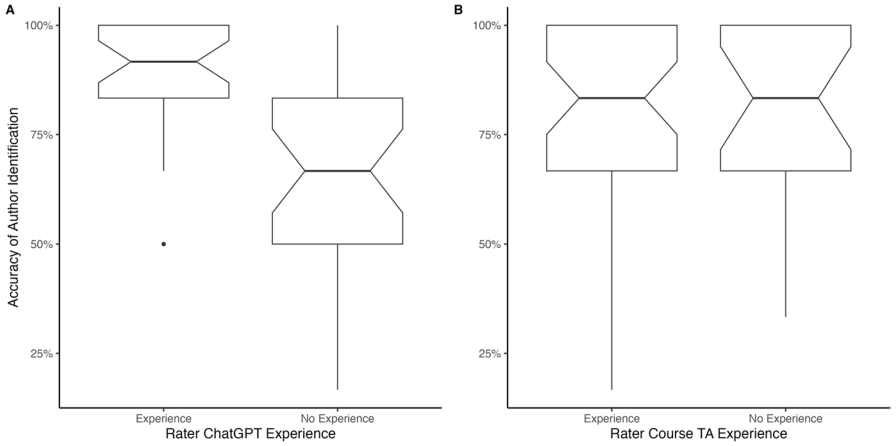

Fig. 3 Author identification accuracy by rater experience with ChatGPT (**A**) and with course content (**B**)

$p < 0.001$), with ChatGPT users achieving 89% accuracy on average (SD = 13%) and teaching assistants with no AI experience achieving 68% accuracy on average (SD = 23%). Previous course experience was not a significant predictor of identification accuracy ($b = 0.06$, $p = 0.85$).

Relationships between Humanlikeness Ratings, Grades, and Open-Ended Rationales

During the rating process, teaching assistant raters were not privy to grades awarded to each response by the second author. Figure 4 displays distributions of student-authored essay grades by humanlikeness rating. While a relatively small number of Student-authored responses were erroneously rated as "Definitely Machine-Generated", these responses were those with better median grades than responses accurately identified as Human Written.

Out of the ten total authors, the ChatGPT "No History" Author yielded the lowest average identification accuracy across raters (M = 58%, SD = 29%). That is, the "No History" ChatGPT responses were the most likely to be erroneously identified as "Human Written". In contrast, essays written by the ChatGPT "With History" Author were much more likely to be accurately identified as "Machine Generated" (M = 83%, SD = 15%). Across raters, Student Author 15's responses were the most likely to be erroneously identified as a "Machine Generated" (accuracy M = 69%, SD = 16%), and Student Author 42's responses were most likely to be accurately identified as "Human Written" (accuracy M = 94%, SD = 9%). Figure 5 displays the rationales provided by raters for their humanlikeness ratings across levels of identification accuracy for ChatGPT (High: "With History" and Low: "No History") and Student authors (High: Student 42 and Low: Student 15). Rationales were coded using the annotation scheme described by Clark and colleagues (2021). Word Choice (references to specific words or phrases that raters deem "human" or "machine"), Level of Detail (references to the complexity of the text), and Genre/

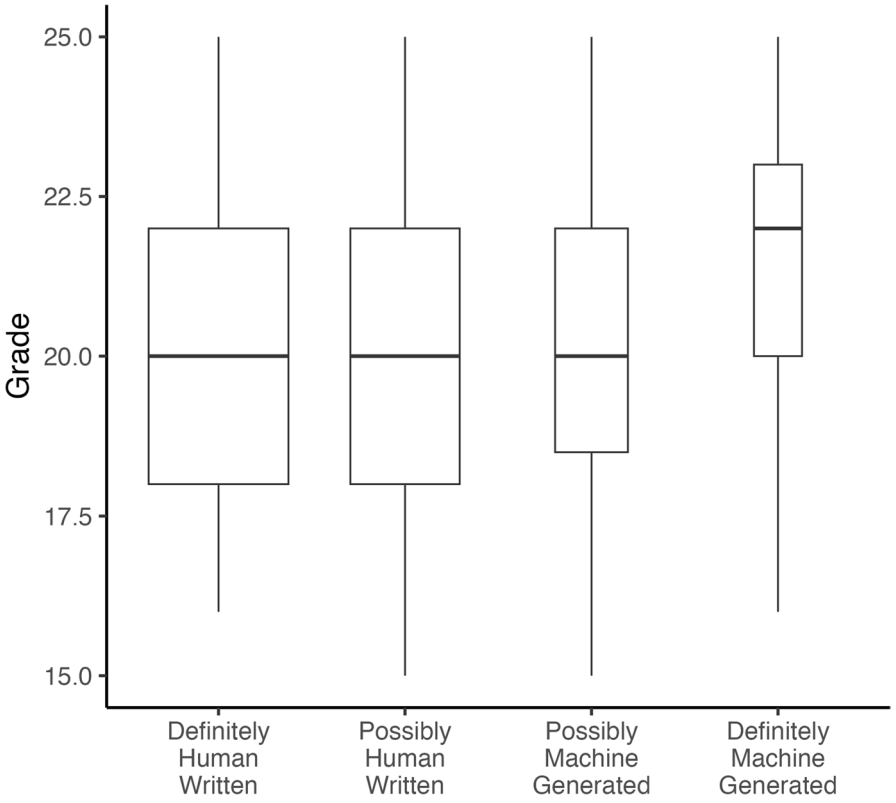

Fig. 4 Distribution of student grades by humanlikeness ratings. Boxplots illustrate grade medians (out of a possible 25) and interquartile ranges for student-authored essays. Boxplot width is proportional to the number of items rated at each level of "humanlikeness"

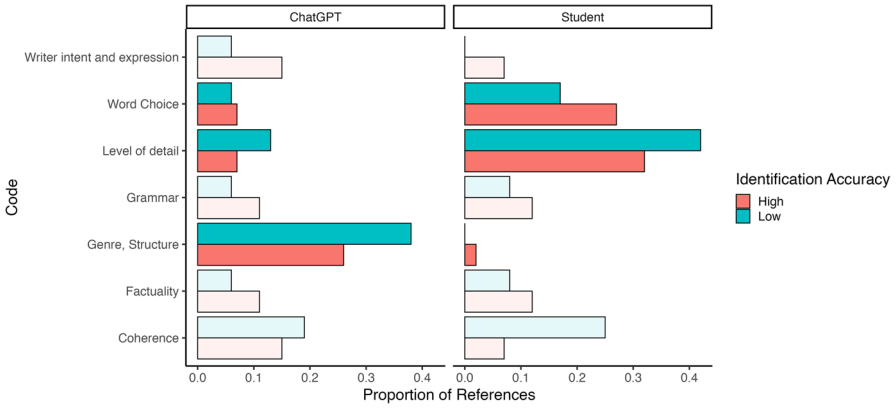

Fig. 5 Rationale coding for Author IDs with the highest and lowest identification accuracy by author. Along the x-axis is the proportion of references to each rationale code (Clark et al., 2021) across all six teaching assistant raters. Code categories that most discriminated between ChatGPT and student authors are highlighted with a darker fill

Structure (references to the genre and structure of the text and its adherence to reader expectations and style norms) emerged as best able to discriminate between ChatGPT and Student authors, irrespective of raters' identification accuracy.

Genre/Structure was the most-referenced rationale code for ChatGPT responses. Teaching assistants who accurately identified the ChatGPT "With History" Author as "Machine Generated" provided rationales related to the response's Genre/Structure including "…repeats parts of the question prompt almost exactly in the answer" and "…the use of explanations with colons feels like AI to me". Teaching assistants who inaccurately identified the ChatGPT "No History" as "Human Written" provided rationales related to the response's Genre/Structure such as "Written in bullet point/list form".

Level of Detail and Word Choice were the most-referenced rationale codes for the two Student-authored responses. Teaching assistants who accurately identified responses by Student Author 42 as being "Human Written" provided rationales related to Word Choice including "…use of the word 'you'" and "wording of 'wants to move' seems like a student" and rationales related to Level of Detail ("I think they may have put extra information to make sure to cover all the exam prompts/questions—which sounds like a person would do!"). Teaching assistants who inaccurately identified responses by Student Author 15 as being "Machine Generated" provided rationales of Word Choice ("I don't see a student using 'roll, pitch, and yaw'") and Level of Detail ("I think some machine generated pieces seem to have a lot of information but missing some details of what the prompt is asking").

Identification of AI Authored Responses

Teaching assistant estimates of the number of AI-authored responses ranged from two to five (the correct number of AI-authored responses was two out of a possible ten). When asked to rank Author IDs by likelihood of being AI, teaching assistants accurately identified an average of 67% of the ChatGPT-authored responses (SD = 26%). Two of the teaching assistants with ChatGPT experience correctly identified both ChatGPT authors, but none of the teaching assistant raters reported that they felt sure enough about their identification of the AI-authored exams to be comfortable reporting such a response as an instance of academic dishonesty in a real-world setting.

Discussion

The current study examined the performance of GPT-3.5 on undergraduate neuroscience course materials, and the ability of doctoral student teaching assistant raters to identify AI-generated text under conditions more similar to a typical classroom environment compared to previous studies. GPT-3.5 performed around the class average or better, in some cases outperforming the vast majority of student test takers (i.e. Exam 1, where both ChatGPT respondents scored better than 79% of student respondents). At the individual essay and quiz level, there was substantial variability

in GPT-3.5's performance, with occasional instances of much poorer performance. Our study adds to the literature characterizing ChatGPT's performance in undergraduate science education and is, to our knowledge, the first to explore ChatGPT's performance on neuroscience course materials. Our results mirror those in other fields within undergraduate science education including physics (Kortemeyer, 2023; West, 2023) and chemistry (Clark et al., 2023; Watts et al., 2023), that have demonstrated impressive but variable performance of ChatGPT on common assessment formats in STEM education. In particular, ChatGPT's strong performance on average in response to introductory neuroscience essay prompts validates concerns about its impact on foundational undergraduate coursework, in which learning objectives are focused on knowledge acquisition at lower levels of Bloom's taxonomy (Newton & Xiromeriti, 2023).

Doctoral student teaching assistants were adept at identifying ChatGPT-generated responses at levels better than chance. Raters' identification accuracy differed across author types, with significantly poorer identification accuracy for the ChatGPT "No History" condition relative to Student- and ChatGPT "With History"-authored essays. While we do not have conclusive evidence to explain this discrepancy, one possibility is that the poorer overall performance (and greater variability in performance across essay topics) of the "No History" condition may have resulted in more variable humanlikeness ratings. Degree of experience with the course (i.e. prior TA experience in the course) did not significantly impact identification accuracy. In contrast, direct experience with ChatGPT was a significant predictor of identification accuracy. Raters with ChatGPT experience achieved high levels of accuracy (> 80%) on average compared to those without. Despite strong identification performance, none of the teaching assistant raters reported being confident enough in their assessment to formally report essays suspected as being machine-generated in a real-world context. This highlights the challenges associated with accurately identifying AI-generated responses and contributes to the ongoing discussion of the humanlikeness of machine-generated text (Clark et al., 2021; Jakesch et al., 2023). Our results provide a number of insights into the potential impacts of ChatGPT on undergraduate education. Anticipating these impacts is important as an increasing number of students adopt AI-enhanced tools, with early evidence suggesting that students tend to continue to use AI-enhanced tools once they have been introduced to them (Noy & Zhang, 2023).

Performance of ChatGPT on Undergraduate Neuroscience Assessments

While ChatGPT's overall performance was strong, instances of subpar responses were notable on some of the essay and quiz items. This discrepancy might be partially attributed to the use of existing course rubrics to grade both the student- and AI-authored essay responses. Grading rubrics were strongly aligned with the design of the course itself. Given the methodological choice to rely on existing rubrics, it is possible that in some cases ChatGPT provided an *accurate* response, but that this response lacked particular details that were emphasized in course lectures and assignments and that were not included in the prompts provided to ChatGPT. For

example, for the Synaptic Transmission response produced in the "No History" condition, the response generated by ChatGPT described the process of synaptic transmission at a higher level of analysis with considerably less detail than expected and outlined in the grading rubric (i.e. ChatGPT provided a statement that arrival of an action potential at the axon terminal "triggers release of neurotransmitter" without explaining the process that triggers this release). This results in a response that is "accurate", but that is underspecified in the context of this particular course.

In other cases, the most likely explanation for ChatGPT's poor performance were frank inaccuracies, sometimes referred to as "hallucinations" (Ji et al., 2023). Inaccuracies contributed considerably to several cases in which ChatGPT performed more poorly than student authors. For example, in response to a Quiz 3 question testing knowledge of anatomical directions, when prompted with "Your chin is _____ to your neck", ChatGPT erroneously chose 'caudal' and even backed up this response with the false statement: "the neck is positioned above or superior to the chin". While this example is particularly stark, in other instances, inaccurate statements were much more subtle. For example, in response to the Synaptic Transmission essay question, ChatGPT provided the following response including a subtle inaccuracy: "This influx or efflux of negative ions hyperpolarizes the postsynaptic membrane…" These examples of errors on introductory neuroscience exams and quizzes confirm that GPT-3.5 is not error proof, even for concepts for which it should have abundant training data. Subtle inaccuracies in particular are likely to be challenging for students to identify when they are still learning foundational material.

Identification of AI-Generated Text in Classroom Contexts

Our study presents a departure from previous AI-identification research, with teaching assistant raters surpassing the accuracy of blinded raters in previous studies (Clark et al., 2021; Jakesch et al., 2023). This success might be attributable to the domain expertise of our raters. While teaching assistants varied in direct experience with this particular course, all raters had experience in the field and relevant domain expertise. This contrasts with previous studies in which raters were drawn from the general population. Our study also differs from previous studies in that ratings were conducted in-person, versus online via crowdworking platforms (e.g. Amazon Mechanical Turk or Lucid; Clark et al., 2021; Jakesch et al., 2023). Our findings align with a study that demonstrated that quality ratings for text generated by GPT-2 and human authors significantly differs when ratings are made by crowdworkers compared to those with domain expertise (Karpinska et al., 2021). Finally, it is possible that identifying machine-generated text is more straightforward in cases where such text is the exception and not the norm. That is, in our study 20% of essay responses were machine-generated in the rating task, whereas previous AI-identification studies included an equal balance of AI- versus human-authored responses (Clark et al., 2021; Jakesch et al., 2023). These distinctions shed light

on the nuanced nature of identifying AI-generated text in real-world educational settings.

In this study, the numeric humanlikeness ratings provided by teaching assistants effectively discriminated between AI- and student-authored essays, surpassing chance-level accuracy. The explicit *rationales* for these ratings, however, were often conflicting. For example, one rater explained a rating of "Human Written" by explaining that the response was "written in bullet point/list form instead of an essay". For the same essay response, a different rater explained their conflicting rating of "Machine Generated" with the rationale: "Answers in a list with colons, combined with the writing style make me lean toward AI". These findings expand upon previous research highlighting the limitations of human judgement in identifying AI-generated text (Clark et al., 2021; Jakesch et al., 2023). Jakesch and colleagues demonstrated that, even when identification accuracy is low, there is often better-than-chance *agreement* among human raters regarding the authorship (AI vs. human) of a given text. This indicates that raters tend to rely on "shared but flawed" heuristics when tasked with identifying AI-generated language (2023).

In their study, Jakesch and colleagues assessed the humanlikeness judgments of raters using an existing annotation scheme (Clark et al., 2021). In addition to the explicit rationales provided by raters, the researchers investigated which linguistic features within the stimulus texts were consistently associated with humanlikeness judgments (AI vs. human). Their findings indicated that certain linguistic features within the stimulus texts could be identified by blinded raters and, remarkably, these features could differentiate authorship better than chance. However, raters who were directly asked to rate humanlikeness had significantly poorer performance. This difference in discrimination accuracy between explicit rationales and "revealed" rationales suggests the presence of several flawed human heuristics when identifying machine-generated text. For example, raters were more inclined to label texts with grammatical errors as AI-generated, even though grammatically flawed texts were more likely to be human-written. Similarly, raters tended to rate texts with long or infrequently used words as AI-generated when they were more likely to be human written. The results of this study highlight the disparities in features between AI- and human-generated text that human authors can recognize; yet, when explicitly asked to determine authorship, human raters frequently exhibit inaccuracies due to flawed heuristics. In our rationale data, we observed similar patterns in the identification of features that differentiated authorship (e.g. word choice and level of detail were mentioned more frequently for student- vs. AI-authored essays); however, recognition of these features did not necessarily result in accurate humanlikeness ratings.

Potential Impacts of AI Tools in Foundational Undergraduate Education

Our study underlines potential pitfalls associated with integrating AI tools into foundational undergraduate education. The course described here has historically taken a "writing to learn" approach, in which students increase their understanding of complex processes by explaining them in writing, soliciting feedback from peers and instructors, and continuing to refine their understanding (Keys, 1999; Klein &

Boscolo, 2016; Finkenstaedt-Quinn et al., 2021). Learning objectives within the course are primarily at the "understanding" level of Bloom's taxonomy (Bloom et al., 1956). Critically, this "lower" level along the Taxonomy should not be taken as a proxy for the challenge the course represents for students new to neuroscience. The concepts and processes under study are complex and challenging to learn. Students' ability to explain neuronal signaling, synaptic transmission, and the structure and function of specialized neural systems set a *foundation* upon which future courses will build. As described by Newton and Xiomeriti, establishing understanding at lower levels of Bloom's taxonomy in foundational coursework is critical for future cultivation of higher-order skills (2023). Our results demonstrate that ChatGPT is particularly adept at providing rapid and well-structured responses to prompts that require processing at the "understanding" level. Given this strong performance, undergraduate students may be tempted to sidestep the challenge inherent in learning foundational but difficult course concepts and rely on rapidly generated and generally high-quality GPT responses, underscoring the heightened potential drawbacks of reliance on AI in foundational course contexts.

This challenge echoes broader debates about "cognitive offloading" and the need to balance AI support with the cultivation of essential cognitive skills. The potential impact of AI-enhanced tools on education has been compared to prior technological advances, including improvements in pocket calculators and the advent of spelling and grammar checkers. On this view, ChatGPT and other generative AI tools facilitate "cognitive offloading", reducing the cognitive demands of a particular task, and freeing up cognitive resources for higher-order cognitive tasks (Risko & Gilbert, 2016). We argue that while cognitive offloading can be a boon for individuals who have already mastered the component skills being "offloaded" to a particular technological tool, it can become problematic when novice learners sidestep component skills entirely. In a course similar to that reported here, that takes a "writing to learn" approach, use of ChatGPT to generate initial drafts has the potential to reduce the utility of this practice.

A helpful framework for considering the possible impacts of using ChatGPT for cognitive offloading in undergraduate courses is outlined by Paas and colleagues and divides cognitive load associated with course assignments and assessments into intrinsic, extraneous, and germane cognitive load (Paas et al., 2003). Intrinsic cognitive load refers to cognitive demands that are intrinsic to the material being learned. In describing intrinsic cognitive load, Paas and colleagues highlight the degree of "interactivity" in the material to be learned. Material that can easily be broken up into discrete, unrelated facts has low intrinsic cognitive load, whereas material in which the various components "interact" and depend on one another, such that they cannot be simplified or stripped of their context are high in intrinsic cognitive load. The neuroscience course described here involves significant "intrinsic" cognitive load; using AI-enhanced tools to reduce this intrinsic cognitive load robs students of important learning opportunities. In contrast, extraneous cognitive load refers to cognitive demands that are unnecessary or unrelated to learning objectives and that interfere with the material to be learned. For example, poorly written assignment instructions might require high working memory demands that are not relevant to the material to be learned, but that nevertheless demand significant cognitive

resources. In cases where producing structured writing is tangential to a particular learning objective, tools like ChatGPT might be usefully employed to reduce extraneous cognitive load, without negatively impacting learning outcomes. Finally, germane cognitive load refers to cognitive demands imposed by an instructor that enhance learning of the material under study. For example, increasing the amount of effort, engagement, and motivation being exerted while learning increases overall cognitive load, but when aligned with learning objectives enhances learning and retention. Use of ChatGPT to bypass an instructors' intended addition of germane cognitive load may also reduce students' learning and retention of foundational course material.

GPT-3.5 performed similarly to the student class average across exam question in the current study, which might suggest that GPT-3.5 output may be a useful study tool for students who are performing below-average in the course. It is important to consider, however, how struggling students' use of ChatGPT and similar LLMs as a study tool may also result in negative impacts as a result of the "illusion of understanding", in which the ease and linguistic quality of responses generated via ChatGPT provide false assurance of a correct response. Dahlkemper and colleagues investigated students' ability to recognize inaccuracies generated by ChatGPT in the context of undergraduate physics (2023). Students with varying levels of physics background were asked to rate the scientific accuracy of four responses to physics questions of varying difficulty, authored either by ChatGPT or by a team of physics professors. Across the group as a whole, students rated ChatGPT's responses as less accurate than the expert-authored responses. Importantly, however, students were most successful at discriminating between ChatGPT- and expert-authored responses for physics questions that were lower in difficulty. For more complex physics questions, students were less able to identify differences in the scientific accuracy of ChatGPT's responses compared to expert responses. Furthermore, students with lower self-assessed physics knowledge were less able to identify differences in scientific accuracy between the ChatGPT- and expert-authored responses, compared to students with higher self-assessed physics knowledge. These results highlight the challenges of AI hallucinations in the context of undergraduate education: for students who are themselves learning foundational material, identifying inaccuracies in the context of strong, plausible-sounding linguistic output is a significant challenge.

Finally, differences in access to AI-enhanced tools raise concerns about equity across students. Here, we tested the performance and ease of identification of output from GPT-3.5. At the time of writing, GPT-3.5 is freely available via OpenAI with a registered account, but it is possible to access GPT-4 via a subscription service. GPT-4 has proven to be more powerful than 3.5 in its ability to accomplish tasks such as text generation, language translation, and data analysis. This difference in performance between the free- and paid-version of the tool raises significant risks of inequities among students using GPT-3.5 compared to those able to pay for GPT-4.

Amidst these concerns, ChatGPT may offer tangible benefits, particularly for students with varying writing proficiency. Noy and Zhang noted significant reduction in task completion time coupled with improved writing quality (2023). Time improvements were similar across all students, but quality improvements were most notable for students at the lower end of the proficiency range (Noy & Zhang, 2023).

It has been proposed that AI hallucinations may also serve as useful teaching tools, especially for students at more advanced levels (Dahlkemper et al., 2023). Proposals for use of AI hallucinations as teaching tools include as assignments asking students to evaluate AI-generated text for errors or as in-class activities in which instructors walk through AI-generated text and explain where concepts are inaccurately described (Dalhkemper et al. 2023; Hensley, 2024; Nadeem et al., 2024; Salamin et al., 2023). Careful scaffolding is essential to maximize the educational utility of analyzing AI hallucinations, in order to prevent challenges like the "illusion of understanding" described above.

Limitations

This study was conducted using responses obtained from GPT-3.5. In the intervening time between stimulus creation and data analysis, GPT-4 was released. It is possible that GPT-4's responses to exam and quiz question prompts would be more accurate or harder for teaching assistant raters to identify as machine-generated. However, at present GPT-3.5 is still the "free tier" of ChatGPT and is likely the most widely available implementation of ChatGPT in use by students. In addition, ChatGPT's responses have been shown to change over time, with decreases in performance noted for a wide range of tasks (Chen et al., 2023).

This study made use of existing essay exam responses which were handwritten. It is possible that some factor other than response quality may have influenced grades and ratings for the ChatGPT responses. We sought to mitigate against this possibility by having two separate research assistants handwrite "for" the two ChatGPT authors. Rationales provided by the teaching assistant raters focused on response quality, content, and form and suggest that we were successful in reducing differences in stimulus materials that were extraneous to the linguistic quality and accuracy of responses.

Finally, the choice to include only two ChatGPT authors in the stimulus set limited our ability to further characterize the rationale data. This decision was meant to increase the ecological validity of our stimulus set and rating process, and results of our analysis of the rationale data for even a small subset of responses aligns with previous findings (Clark et al., 2021; Jakesch et al., 2023).

Conclusions

GPT-3.5 achieved passing grades on introductory neuroscience course materials, often outperforming a substantial proportion of student test takers. Our results suggest that ChatGPT will have significant impacts on common assessment practices in undergraduate science education, particularly for courses that continue to implement online or at-home assessments post-pandemic. Teaching assistant raters were able to identify AI-generated text with better than chance accuracy but reported that they did not feel confident enough in their determinations to initiate academic integrity processes in real-world settings. Identification accuracy was significantly impacted

by hands-on experience with ChatGPT, suggesting that instructors who wish to lessen potential negative impacts of this emerging technology on their courses should engage with it directly, in order to develop a sense for its capabilities, flaws, and characteristic output. ChatGPT and other LLMs pose a particular challenge to foundational coursework. As the educational landscape continues to integrate AI-enhanced tools, it is imperative to address concerns about foundational skill development, the potential for negative impacts of cognitive offloading, and issues of equity related to tool accessibility. Our findings contribute to the ongoing dialogue on the role of AI in education and underscore the need for nuanced approaches to maximize benefits while mitigating pitfalls.

Appendix

Table 4

Table 4 Accuracy ~ Author + AI Experience + Course Experience + (1|item) + (1|rater), family = binomial

Fixed Effects	*Est*	*SE*	*z value*	*p-value*
Intercept	**0.51**	**0.24**	**2.09**	**0.04**
Author Contrast 1 [Student = -1 vs. ChatGPT "With History" = 1]	0.10	0.24	0.41	0.68
Author Contrast 2 [Student/ChatGPT "With History" = -1 vs. ChatGPT "No History" = 2]	**-0.43**	**0.15**	**-2.97**	**0.003**
AI Experience [reference group = No Experience]	**1.44**	**0.30**	**4.79**	**< 0.001**
Course Experience [reference group = No Experience]	0.06	0.29	0.19	0.85
Random Effects	*Variance*	*SD*		
Item (Intercept)	0	0		
Rater (Intercept)	0	0		
Observations: 360; Items: 60; Raters: 6				

The dependent measure is a binary measure of accuracy (1 = accurate identification, 0 = inaccurate identification). Statistically significant effects are highlighted in **bold**

To compare identification accuracy across essay Author types, levels of Author were coded using Reverse Helmert coding. Reverse Helmert coding compares each level of a categorical factor to the mean of previous levels. The first Author Contrast coefficient compares Student authored essays to ChatGPT "With History" authored essays. The second Author Contrast coefficient compares the ChatGPT "No History" authored essays against the average of Student- and ChatGPT With History" authored essays

Note that there was little variation across clusters of items and raters. These random effects have been maintained here given the nested structure of the data, but all significant results and fixed effect estimates are stable in a logistic effects model without random effects and the same fixed effect structure

Acknowledgements Thank you to Elizabeth Ancel, Calvin Duong, Miriam Kornelis, HaeJi Lee, Emma Morseth, Michael Smith, and Dana Urbanski for their help in stimulus development and response ratings.

Author's Contribution NVC: Conceptualization, Methodology, Formal Analysis, Investigation, Resources, Writing—Original Draft, Writing—Review & Editing, Visualization, Supervision, Project Administration. OV: Formal Analysis, Investigation, Data Curation, Writing—Review & Editing, Project Administration.

Funding This study was unfunded.

Data Availability The data that support the findings of this study are available from the corresponding author upon request.

Declarations

Ethics Approval This study was reviewed and deemed exempt by the University of Minnesota Institutional Review Board.

Competing Interests The authors have no relevant financial interests to disclose. Both authors serve in teaching-related roles in the course from which study materials were drawn. The study was conducted using de-identified secondary data after the course had fully concluded.

References

Alexander, K., Savvidou, C., & Alexander, C. (2023). Who wrote this essay? Detecting AI-Generated writing in second language education in higher education. *Teaching English with Technology, 23*(2), 25–43.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (Eds.). (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain*. David McKay.

Bordt, S., & von Luxburg, U. (2023). ChatGPT participates in a computer science exam. *arXiv preprint arXiv:2303.09461*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Buchanan, J., Hill, S., & Shapoval, O. (2024). ChatGPT hallucinates non-existent citations: evidence from economics. *The American Economist, 69*(1), 80–87.

Chan, C. K. Y. (2022). A review of the changes in higher education assessment and grading policy during covid-19. *Assessment & Evaluation in Higher Education, 48*(6), 874–887.

Chen, L., Zaharia, M., & Zou, J. (2023). How is ChatGPT's behavior changing over time?. *arXiv preprint arXiv:2307.09009*.

Choi, J. H., Hickman, K. E., Monahan, A. B., & Schwarcz, D. (2022). ChatGPT goes to law school. *Journal of Legal Education, 71*, 387.

Clark, T. M., Anderson, E., Dickson-Karn, N. M., Soltanirad, C., & Tafini, N. (2023). Comparing the Performance of College Chemistry Students with ChatGPT for Calculations Involving Acids and Bases. *Journal of Chemical Education, 100*(10), 3934–3944.

Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., & Smith, N. A. (2021). *All That's "Human" Is Not Gold: Evaluating Human Evaluation of Generated Text* (arXiv:2107.00061). arXiv. http://arxiv.org/abs/2107.00061 https://doi.org/10.1371/journal.pone.0025085

Fergus, S., Botha, M., & Ostovar, M. (2023). Evaluating academic answers generated using ChatGPT. *Journal of Chemical Education, 100*(4), 1672–1675.

Finkenstaedt-Quinn, S. A., Petterson, M., Gere, A., & Shultz, G. (2021). Praxis of Writing-to-Learn: A Model for the Design and Propagation of Writing-to-Learn in STEM. *Journal of Chemical Education, 98*(5), 1548–1555.

Hensley, R. (2024). An AI Workshop for the Overwhelmed and Uninterested. *Teaching and Generative AI*.

Holland, A., & Ciachir, C. (2024). A qualitative study of students' lived experience and perceptions of using ChatGPT: Immediacy, equity and integrity. *Interactive Learning Environments,* 1–12.

Ippolito, D., Duckworth, D., Callison-Burch, C., & Eck, D. (2020). *Automatic Detection of Generated Text is Easiest when Humans are Fooled* (arXiv:1911.00650). arXiv. http://arxiv.org/abs/1911.00650

Jakesch, M., Hancock, J. T., & Naaman, M. (2023). Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences, 120*(11), e2208839120.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), 1–38.

Karpinska, M., Akoury, N., & Iyyer, M. (2021). The perils of using mechanical turk to evaluate open-ended text generation. *arXiv preprint arXiv:2109.06835*.

Kerrigan, J., Cochran, G., Tabanli, S., Charnley, M., & Mulvey, S. (2022). Post-covid changes to assessment practices: A case study of undergraduate STEM recitations. *Journal of Educational Technology Systems, 51*(2), 192–201.

Keys, C. W. (1999). Revitalizing instruction in scientific genres: Connecting knowledge production with writing to learn in science. *Science Education, 83*(2), 115–130.

Khalil, M., & Er, E. (2023). Will ChatGPT Get You Caught? Rethinking of Plagiarism Detection. In P. Zaphiris & A. Ioannou (Eds.), *Learning and Collaboration Technologies. HCII 2023. Lecture Notes in Computer Science.* (Vol. 14040). Springer, Cham.

Klein, P. D., & Boscolo, P. (2016). Trends in research on writing as a learning activity. *Journal of Writing Research, 7*(3), 311–350.

Kortemeyer, G. (2023). Could an artificial-intelligence agent pass an introductory physics course? *Physical Review Physics Education Research, 19*(1), 010132.

Malik, A., Khan, M. L., & Hussain, K. (2023). How is ChatGPT transforming academia? *Examining its impact on teaching, research, assessment, and learning*. https://doi.org/10.2139/ssrn.4413516

Mamo, Y., Crompton, H., Burke, D., & Nickel, C. (2024). Higher education faculty perceptions of ChatGPT and the influencing factors: a sentiment analysis of X. *TechTrends, 68*(3), 520–534.

Newton, P. M., & Xiromeriti, M. (2023). ChatGPT Performance on MCQ Exams in Higher Education. A Pragmatic Scoping Review. *Assessment & Evaluation in Higher Education,* 1–18.

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science, 381*(6654), 187–192.

OpenAI. (2023). GPT-4 Technical Report.

Parker, L., Carter, C., Karakas, A., Loper, A. J., & Sokkar, A. (2024). Graduate instructors navigating the AI frontier: The role of ChatGPT in higher education. *Computers and Education Open, 6*, 100166.

Perkins, M. (2023). Academic Integrity considerations of AI Large Language Models in the post-pandemic era: ChatGPT and beyond. *Journal of University Teaching & Learning Practice, 20*(2), 07.

Pursnani, V., Sermet, Y., Kurt, M., & Demir, I. (2023). Performance of ChatGPT on the US fundamentals of engineering exam: Comprehensive assessment of proficiency and potential implications for professional environmental engineering practice. *Computers and Education: Artificial Intelligence, 5*, 100183.

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences, 20*(9), 676–688.

Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching, 6*(1), 342–363.

Salamin, A. D., Russo, D., & Rueger, D. (2023). ChatGPT, an excellent liar: how conversational agent hallucinations impact learning and teaching. In *Proceedings of the 7th International Conference on Teaching, Learning and Education*.

Villasenor, J. (2023). How ChatGPT Can Improve Education, Not Threaten It. Scientific American. https://www.scientificamerican.com/article/how-chatgpt-can-improve-education-not-threaten-it/

Walters, W. H. (2023). The effectiveness of software designed to detect AI-Generated writing: a comparison of 16 AI text detectors. *Open Information Science, 7*(1), 20220158.

Walters, W. H., & Wilder, E. I. (2023). Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports, 13*(1), 14045.

Watts, F. M., Dood, A. J., Shultz, G. V., & Rodriguez, J. M. G. (2023). Comparing Student and Generative Artificial Intelligence Chatbot Responses to Organic Chemistry Writing-to-Learn Assignments. *Journal of Chemical Education, 100*(10), 3806–3817.

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

West, C. G. (2023). AI and the FCI: Can ChatGPT project an understanding of introductory physics?. *arXiv preprint arXiv:2303.01067*.

Woo, D. J., Guo, K., & Susanto, H. (2023). Cases of EFL Secondary Students' Prompt Engineering Pathways to Complete a Writing Task with ChatGPT. *arXiv preprint arXiv:2307.05493*.

Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B., & Yang, Q. (2023). Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1–21).

Authors and Affiliations

Natalie V. Covington[1,2] · **Olivia Vruwink**[1]

✉ Natalie V. Covington
nvcoving@umn.edu

[1] Department of Speech-Language-Hearing Sciences, University of Minnesota, Minneapolis, MN, USA

[2] Courage Kenny Rehabilitation Institute, Allina Health, Minneapolis, MN, USA