When Chatting Isn't Cheating: Mining and Evaluating Student Use of Chatbots and Other Resources During Open-Internet Exams

David A. Joyner
College of Computing
Georgia Institute of Technology
david.joyner@gatech.edu

Zoey Anne Beda
College of Computing
Georgia Institute of Technology
zbeda3@gatech.edu

Michael Cohen
College of Computing
Georgia Institute of Technology
mcohen66@gatech.edu

Melanie Duffin
College of Computing
Georgia Institute of Technology
mduffin@gatech.edu

Amy Garcia Fernandez
College of Computing
Georgia Institute of Technology
afernandez98@gatech.edu

Liz Hayes-Golding
College of Computing
Georgia Institute of Technology
ehayesgolding3@gatech.edu

Jonathan Hildreth
College of Computing
Georgia Institute of Technology
jhildreth9@gatech.edu

Alex Houk
College of Computing
Georgia Institute of Technology
ahouk3@gatech.edu

Rebecca Johnson
College of Computing
Georgia Institute of Technology
rebeccaj@gatech.edu

Kayla Matcheck
College of Computing
Georgia Institute of Technology
kgrotsky3@gatech.edu

Ana Santos
College of Computing
Georgia Institute of Technology
asantos61@gatech.edu

ABSTRACT

This study examines log data from proctored examinations from two classes offered as part of a large online graduate program in computer science. In these two classes, students are permitted to access any internet content during their exams, which themselves have remained largely unchanged over the last several semesters. As a result, when ChatGPT and other more sophisticated chatbots arrived in 2022, students were permitted to begin using these tools during their exams. Proctoring tools used during these examinations capture what internet resources are used. This study mines these data regarding what resources use during examinations and evaluates whether access to more sophisticated AI tools has had a notable impact on student performance, as well as how they use these tools. This study also examines what other resources students access, providing insights into the need for localization and accessibility technologies. Ultimately, this study finds that there is at present no strong data to indicate that using AI during these examinations has improved student performance: grades among students who use AI are approximately the same as those among students who do not, and the overall class average on these tests has not changed since the pre-ChatGPT era.

Keywords

ChatGPT, exams, chatbots

1. INTRODUCTION

The rapid rise—at least from the perspective of the general public—of generative AI tools like ChatGPT has left many educators in a quandary over the best way to respond to the emergence of these new tools. Some have been quick to generally categorize undisclosed use of generative AI as a form of plagiarism or academic misconduct [9, 39]. Others have taken a more measured approach, acknowledging the risks that generative AI poses while still exploring ways to use it as a productive tool in education [4, 51, 66]. Embedded in this discourse is an implicit question: on an imaginary spectrum from contract cheating [18, 25, 26, 46] to calculators [29, 41], where does generative AI lie [34, 69]?

There are numerous challenges to answering this question. While contract cheating and the use of calculators in mathematics both had decades of time to develop, generative AI emerged rapidly, building on the existing ubiquity of the internet and smart devices to immediately land in the hands of billions of people worldwide. As such, educators have not had the long timescale to carefully examine how students

D. Joyner, Z. A. Beda, M. Cohen, M. Duffin, A. G. Fernandez, L. Hayes-Golding, J. Hildreth, A. Houk, R. Johnson, K. Matchek, and A. Santos. When chatting isn't cheating: Mining and evaluating student use of chatbots and other resources during open-internet exams. In B. Paaßen and C. D. Epp, editors, *Proceedings of the 17th International Conference on Educational Data Mining*, pages 143–156, Atlanta, Georgia, USA, July 2024. International Educational Data Mining Society.

https://doi.org/10.5281/zenodo.12729788

use these new tools and to thoughtfully assess whether they are being used as a partner for learning or as a substitute for it. Understanding the way in which students—absent any strong guidance or limitations—use generative AI tools in educational contexts is crucial to informing policies and instructions on effective use.

This investigation must be conducted across numerous different levels, subjects, and types of assessment, as generative AI's potential differs from context to context. Toward that end, in this work we investigate its use in one such context: during exams in two graduate-level computer science courses. Exams in these courses are notable because students are permitted to use any resources *except* for interacting with other people while completing the exam; as such, they are allowed to use ChatGPT and other conversational and generative AI tools *during* the exam. At the same time, the exams' proctoring tools can keep a log of what resources students access during the exams, allowing teachers and researchers to investigate how and to what extent students are using these tools when permitted to do so.

In this study, we mine the resources students access during these exams. Using this dataset, we perform three studies.

1. First, we quantitatively evaluate the extent to which students use generative AI tools on these exams and connect these usage statistics to exam performance data. Understanding whether students are accessing a major advantage by leveraging AI tools may provide crucial insights into their acceptability in education.

2. Second, we qualitatively investigate *how* students use these generative AI tools: are they using them to confirm their existing answers? To fill in the gaps on areas on which they are unsure? To complete the exam in its entirety with little to no knowledge of their own?

3. Third, because of the easy access to this data granted by this initial mining exercise, we further investigate what other resources students often use during these exams. Although tangential to the core question of AI usage, this investigation provides some additional general insights, such as on the relative importance of translation software for students for whom English is not their first language.

2. RELATED WORK

This work exists at the intersection of three broad trends in digital education: digitally proctored examinations, human-AI collaboration, and AI in education. We give background to each of these three below.

2.1 Digital Proctoring

Ensuring academic integrity has long been a challenge for distance education. Prior to the rise of digital and online education, testing centers were regularly used where students would go to a central location where a human proctor would deliver the tests to students and ensure the student was abiding by exam rules [27]. The rise of the internet drove an increased interest in distance education as new technologies made delivering instruction and assessment more scalable and feasible. However, concerns over integrity remained, motivating the rise of digital proctoring efforts.

Digital proctoring in general uses technologies already owned by the typical online learner—a webcam, a microphone, a computer capable of screen capture, and access to high-speed internet—to remotely monitor the student's testing environment and assess whether they are following test rules. Early efforts toward digital proctoring essentially recreated the live testing experience by having a human proctor watch remotely in realtime [37, 61]; this approach was also common during the early days of the COVID-19 pandemic [53, 43, 54]. This arrangement, however, required students to schedule their exams in advance and required significant advance notice to arrange adequate proctoring support, presenting a challenge to scale. Additionally, faculty and integrity officers were forced to rely on written reports from human proctors unaffiliated with their institution without any reviewable evidence.

These weaknesses gave rise to what is now recognized as the more common form of digital proctoring. Most modern digital proctoring tools—like Honorlock, Proctortrack, ProctorU, and Examity—use students' own devices to record their exam sessions, often including capturing the student's face and voice via their webcam and their screen via screen-capture [50, 5, 6]. These recordings are then funneled through computer vision algorithms and other AI models to automatically flag instances of suspicious behavior. These instances are then reviewed by humans—either employed by the proctoring company or associated with the class—to determine which represent legitimate misconduct.

The invasiveness of digital proctoring has sparked a push-back from students, faculty, administrators, and politicians alike [8, 19, 28, 40, 45, 65, 67]. Others have referred to digital proctoring as a necessary evil [42, 64] and even found that some students appreciate these tools' role in safeguarding the value of their degree [14]. Some even note the role these tools can play in improving pedagogy: they can allow teachers to use synchronous, collocated classroom time for more valuable learning activities than using this time on independent assessment [7], or they could be used as the foundation of systems for reacting to students' affect to tailor the instructional experience [12, 57, 71].

The question of when and whether digital proctoring is appropriate in a given situation is a complicated question involving numerous trade-offs. However, one advantage it can give is allowing teachers to develop a greater understanding of how their students approach tests. Access to digital proctoring may give teachers the confidence to allow students to use resources or engage in behaviors they fear might be misused because they can monitor those behaviors and adjust accordingly; without this safeguard, they may simply prohibit those behaviors from the outset. This research builds on a similar benefit of access to digital proctoring data: while the default reaction to the rise of generative AI might be to get rid of open-book tests and access to web-based materials during assessments, digital proctoring can allow teachers and researchers to understand how students *are* using these tools prior to making any strong decisions.

2.2 Human-AI Collaboration

By allowing access to AI assistance during exams, tests in these classes are connecting to the rich and growing literature on human-AI collaboration. Students using AI assistants while taking these tests can be seen in some ways as members of a distributed system comprising themselves and an AI agent, and the test pivots to being an assessment of that entire system's ability to demonstrate knowledge [10, 17, 30, 49, 72]. In the context of education, the crucial question here becomes what we expect the system comprised of the student and the AI as a whole to be able to accomplish, and what we expect the student to be able to accomplish independently.

Answers to this question are ever-evolving. After all, today it would be unusual to propose prohibiting students in advanced math classes from using calculators or to students in upper-level writing classes from using word processors and spellcheck. Assessment and technology co-evolve over time, and we may expect to see the question of student collaboration with AI follow the same trajectory [34]. Toward that end, we can borrow from the rich literature on human-AI collaboration to begin to form ideas of what productive collaboration looks like and how to assess it.

2.3 AI in Education

This broader take on the role of AI in education ties into other recent developments in the field, many spawned by the sudden arrival of ChatGPT and similar conversational AI frameworks. The release of these tools sparked a firestorm of rapid research on the role of these new tools in learning [1, 16, 24, 44, 47, 48, 56, 58].

Aside from these broad meta-analysis, position papers, and thought experiments, some of the more nuanced investigations have sought to understand how tools like ChatGPT can be used in particular tasks or for particular purposes. Wang et al. [70] looks to see if ChatGPT can automatically and scalably annotate data for reserach purposes or for teacher development. Phung et al. [55] looks at how these tools can generate feedback on programming syntax. Other work has looked at how tools like ChatGPT can act as a tutor [3, 21, 59]. These research directions in AI in education are perhaps closest to what we are investigating here: how tools like ChatGPT can act as an assistant to students in achieving their goals [62].

3. RESEARCH CONTEXT

As noted in the introduction to this paper, analyses of the use of ChatGPT and similar tools in education must take place in many different contexts, including at different levels, in different fields, and on different types of assessments. This particular research closely examines one such context: open-internet proctored multiple choice tests in two online graduate-level computer science classes.

This context is notable because students enrolling in this program are typically quite technically-savvy: one of the classes under investigation here is itself an artificial intelligence class, and by virtue of entering the program students have demonstrated a high technical aptitude and implicit interest in these emerging technologies. At the same time, research has found that students in programs like these—adult learners needing the flexibility to complete studies alongside professional and family obligations—tend to be primarily motivated by learning itself rather than by external factors [13], and so they may be less likely to engage in using AI to replace rather than augment learning. At the same time, these students tend to be balancing a wider variety of commitments, which may more quickly lead to desperate reliance on AI assistance [2, 11, 60].

3.1 Program Context

The program in which these two courses are offered is a graduate-level program in computer science offered entirely online and asynchronously. The university offering this program follows a semester system, with three semesters—Spring, Summer, and Fall—each year. The Summer semester is 12 weeks long, while Spring and Fall are each 17 weeks. A handful of details of the program are relevant to this study:

- Affordability: The program is priced at approximately $6500 for the entire degree, $540 per class plus student fees each semester. That, coupled with the program's remote and asynchronous nature, means most students do not have as high a financial investment in their success than students in traditional programs.
- Inclusivity: The program accepts any student that it feels has a chance of succeeding, and errs on the side of including students who have the potential to graduate. As such, the program has a higher number of students with weaker prior knowledge than traditional, selective programs. This factor is by design, but may affect the study's observations.
- Size: The program is very large—13,330 students enrolled in Spring 2024. In addition to leading to a greater variety of backgrounds and levels of prior preparation, this size also means that insider information tends to travel fast. Suggestions for using tools effectively (or at times, circumventing rules without getting caught) tend to spread among students quickly.
- Online and Asynchronous: As an online, asynchronous program, students are accustomed to having access to sophisticated tools—including new AI tools—while studying and completing assessments. This greater comfort with these tools likely leads to an increased willingness to try them out during high-stakes timed assessments.

There are other notable details about the program, but these four factors intersect most strongly with this study's research questions. This study examines students with higher than average technical aptitude and comfort using these tools, operating in a program with less financial risk attached to failure, and structured in a way that allows advice on tool usage to propagate more quickly across students. More background on prior work in the context of this and similar programs can be found in prior literature, including the structure of these programs' assessments [15, 20], the nature of these programs' processes and workflows [31, 32, 33, 38, 68], these programs' historical approach to academic integrity [25], and the role of these programs in increasing access to education [22, 23, 35, 36, 52].

Table 1: Enrollment per class during this study

	Spring 2023	Summer 2023	Fall 2023
Class 1	554	443	813
Class 2	457	276	604

3.2 Course Context

This study examines proctored exams within two classes. These two classes are taught by the same instructor, and their exams are structured similarly: each exam contains some number of multiple choice questions, each with multiple right answers. Students are graded on how many answers they correctly mark as well as correctly leave unmarked, effectively pivoting the exam into being graded as a series of 110 to 150 true/false questions.

The two classes—henceforth referred to as Class 1 and Class 2—are rather large each semester. Table 1 shows enrollment in these classes during the time period covered by this study. These enrollment numbers are relatively consistent with earlier semesters as well.

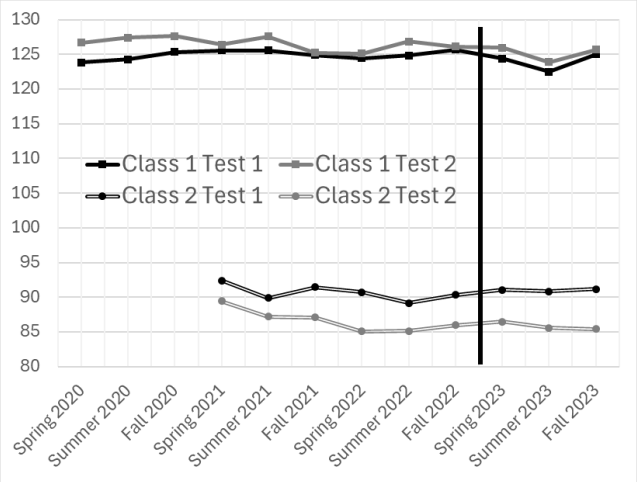
In each class, students complete two proctored tests. Tests in Class 1 are each worth 15% of students' grades; tests in Class 2 are each worth 10%. Tests in Class 1 consist of 30 multiple choice questions each with 5 options, graded out of 150 total possible points; tests in Class 2 consist of 22 questions each with 5 options, graded out of 110 total possible points. Students can launch each test at any time that it is open; once launched, they have two hours to complete tests in Class 1 and 90 minutes to complete tests in Class 2. With the exception of the last three questions on Test 2 in Class 1 (which are updated each year to reflect more recent readings), tests in these classes have remained unchanged over the last several semesters, allowing us to analyze these data as a sort of quasi-experiment. Figure 1 shows the grade trajectory over time; ChatGPT was released while Test 2 of Fall 2022 was open, and the vertical line indicates grades released since ChatGPT's widespread availability. The data does not indicate a systematic increase in scores since ChatGPT became available.

Each class generally leaves its tests open all semester long, though the vast majority of students take the test within one week of the deadline. Answer keys are not shared after grades are calculated; instead, students are given a curated list of feedback based on the questions they got wrong. While the exams allow students to access any web-based resources on the device on which they are completing the exam, a handful of features are disabled: students cannot copy and paste, take screenshots, or print exam content, and they are prohibited from using any device other than the one on which they are completing the exam. These rules provide us some confidence that any use of generative AI during the exams would be captured by the proctoring tool.

3.3 Initial Data Mining

The proctoring system that the program uses has a feature to capture what internet resources are accessed even when students are permitted to access them. This feature was enabled for a subset of the semesters under analysis in this study. Once captured, the proctoring tool presents reviewers with a dashboard of exam sessions, each including a video of the student completing the exam, a screen recording of their exam session, and a list of "flags" associated with the session. "External Resource Accessed" is one such flag; for each "External Resource Accessed" flag, a URL of the resource is provided along with a timestamp of when the resource was accessed.

Figure 1: Average exam grades over time for the two classes and two tests. Class 1's tests (the top lines) are graded out of 150 points; Class 2's tests are graded out of 110. The black vertical line indicates when ChatGPT was released to the public.



To begin this analysis, we constructed a data scraper in Python that accesses a list of exam sessions, and then from each exam session, mines the URLs that were accessed during that exam session. These data are then compiled into a JSON file connecting a user ID for each exam session to a list of URLs accessed during that session. These user IDs could then be connected to gradebook data for to search for connections between resource usage and exam performance.

4. ANALYSIS 1: RATE OF AI USAGE DURING EXAMS

Analysis 1 seeks to understand the extent to which students use tools like ChatGPT during exams, as well as the extent to which such usage appears to have a significant impact on their performance. This impact could take on multiple forms: it might be the case that students who use ChatGPT outperform those who do not, while it could also be the case that students who use ChatGPT underperform relative to their classmates—not because ChatGPT harms their performance, but because underperforming students may be more likely to use it anyway.

4.1 Methodology

The initial cataloged data provide information about each exam session including the scores and the list of domain URLs accessed by each student. These data served as input to a script used to classify each of the domains as *AI* or *not AI*. The following domains were identified for AI usage by

Table 2: AI use per class and exam during this study

		Summer 2023	Fall 2023
Class 1	Exam 1	17.53%	
	Exam 2	22.81%	
Class 2	Exam 1	6.4%	21.3%
	Exam 2	7.2%	33.9%

students: openai, bard, jasper, perplexity, koala, consensus, semanticscholar, claude.ai, hellovaia, chatpdf, pdf.ai, deepai, and botpenguin. Based on these domains, we assign a new column to the data with a boolean variable to indicate if the student had accessed a generative AI tool (True) or not (False). Classifying the domains and establishing the usage of AI for each student allow us to perform multiple analyses to compare the AI usage with the scores data. Analysis 1 is divided into two sub-analyses.

The first sub-analysis focuses on comparing two groups of students: *those who used AI* versus *those who did not use AI* on each exam. We calculate the fraction of students who used AI on each exam to determine how the usage changed as the courses progressed. For each group, we also calculate metrics such as the average, minimum, maximum, and median scores for each exam.

The second sub-analysis focuses on comparing the performance of individual students who had used AI tools on some exams but not on others. In this sub-analysis, we aim to compare each student's performance to themselves, examining if there was a difference in their score on average. We exclude those who never used AI during any exam and those who used AI during all the exams.

For this second sub-analysis, we select students who *used AI on exam 1 but not on exam 2* and those who *used AI on exam 2 but not on exam 1*. Then, the scores of these students are grouped into two categories: *scores when AI was used* versus *scores when AI was not used*. We compare their results irrespective and also respective of the exam. The goal is to evaluate whether there is, on average, a difference in exam performance when students use AI versus when they do not.

4.2 Results

Table 2 shows the percentage of students accessing at least one AI resource on an exam for Class 1 in Summer 2023, as well as Class 2 in both Summer and Fall 2023. In Class 1, the AI usage increased with statistical significance between exam 1 and exam 2 during the Summer session (z = -1.823, p = 0.034 with sample sizes of 389 on exam 1 and 378 on exam 2). In Class 2, the AI usage was particularly low during the Summer session and a notable increase can be seen for the Fall session. AI usage also increased between exam 1 and exam 2 in both terms, although only the increase for Fall 2023 was statistically significant (z = -4.898, $p < 0.001$).

4.2.1 Class 1 Results

Table 3 shows the grade averages across both classes and both exams, as well as across both semesters for Class 2 where exam resource detection was enabled in Fall 2023. These grades are further split between students who did and did not use AI during their exam session.

For sub-analysis 1, the median scores for both exams in Class 1 for Summer 2023 were nearly the same between students using AI and those not using AI. While the averages are close, we observe slight differences in the surrounding quartiles: those who did not use AI see a broader range of scores, as well as a larger interquartile range. We summarize this by observing that the range of grades was narrower for students using AI, but this narrower range was due to both a higher minimum (and higher first quartile score) and a lower maximum (and lower third quartile score); there was no notable impact on the median grade.

For sub-analysis 2, we observed that when individual students used AI on one exam but not the other, irrespective of the order of AI usage across the exams, the average score remained nearly the same. Out of the students in Class 1, 92 were identified as using AI on one exam but not the other. The average score for these students on the exam where AI was used was 82.6%, compared to 81.6% where AI was not used. Additionally, there does not seem to be a significant difference in the minimum (AI Used: 67.3%, No AI Used: 65.3%) and maximum (AI Used: 98.0%, No AI Used: 96.0%) scores. On average, the exam performance of individual students remained consistent when they used AI for one exam but not the other.

We also analyzed the impact of AI usage on exam performance, respective to each exam. 56 students did not use AI on the first exam but used it on the second exam. These students had average scores of 83.4% on exam 1 and 81.7% on exam 2. These averages remain consistent with the class averages (exam 1: 81.6%, exam 2: 82.7%). This suggests that AI usage on the second exam did not significantly affect their performance. Similarly, the 36 students who used AI on the first exam but did not use it on the second exam also showed minimal change in the average scores (exam 1: 81.3%, exam 2: 81.4%). This indicates that not using AI on the second exam had little impact on their performance.

4.2.2 Class 2 Results

Table 3 also shows the statistics for Class 2 across both exams and both Summer and Fall semesters, categorized by students who used at least one AI resource on the exam and students who did not use any AI resources.

For the most part, Class 2 follows the same overall pattern as Class 1 across both exams and both semesters: on most exams, the average score between students who use AI and students who do not is approximately the same. The range, however, is again far wider for students who do not use AI: on both exams in both semesters, students who do not use AI have a lower minimum grade and a higher maximum grade than those who do.

Class 2 is different from Class 1 in one way, however: the grade difference between students who did and did not use AI on exam 2 is notably higher. The difference is not statistically significant, though it is interesting that while the pairwise difference for all other exams ranges from 0.7% to 3.2%, exam 2 in Class 2 has two significantly higher values at 4.5% in Summer 2023 and 3.6% in Fall 2023. These results can be seen in detail in Table 3 and Figure 2.

Table 3: Minimum, first quartile, median, third quartile, and maximum scores for each semester and class, divided between students who did and did not use AI assistance. These data are further visualized in Figure 2.

		Min	Q1	Median	Q3	Max
Class 1, Exam 1						
Summer 2023	AI Used	62.7%	77.3%	81.3%	86.5%	92.0%
	No AI Used	48.7%	76.2%	82.7%	88.0%	98.7%
Class 1, Exam 2						
Summer 2023	AI Used	62.7%	76.7%	83.7%	88.8%	98.0%
	No AI Used	49.3%	77.3%	83.7%	88.7%	97.3%
Class 2, Exam 1						
Summer 2023	AI Used	70.0%	81.8%	83.6%	88.2%	92.7%
	No AI Used	59.1%	79.5%	84.5%	89.1%	99.1%
Fall 2023	AI Used	65.5%	78.9%	83.2%	87.5%	93.6%
	No AI Used	55.5%	80.0%	86.4%	89.1%	96.4%
Class 2, Exam 2						
Summer 2023	AI Used	65.5%	72.3%	78.2%	81.4%	83.6%
	No AI Used	49.1%	77.3%	82.7%	86.4%	95.5%
Fall 2023	AI Used	52.7%	73.6%	79.1%	84.5%	92.7%
	No AI Used	46.4%	77.5%	82.7%	88.2%	96.4%

For sub-analysis 2 within Class 2, students who used AI on one exam but not the other, irrespective of the exam, had similar overall exam averages across the two exams. For Summer 2023, the overall average of the exams where AI was used was 78.3% and the overall average of the exam where AI was not used was 80.5%. For Fall 2023, the overall average of the exams where AI had been used was 79.7% and the overall average of the exam where AI had not been used was 80.9%.

We also observe in Class 2 that students who did not use AI on the first exam but used it on the second exam had lower average scores on exam 1 compared to the overall class average for exam 1. Their average scores for exam 2 also remained lower than the overall class average. The usage of AI on the second exam did not seem to improve the average scores. The group of students who used AI on exam 1 but not on exam 2 had an average exam 1 score slightly below the overall class average, but their average scores for Exam 2, where AI was not used, were closer to the overall class average. The lack of usage of AI on the second exam did not seem to hinder the average scores.

4.3 Discussion

Taken as a whole, these data suggest that using AI did not provide an advantage on these tests. Average scores were roughly the same between AI-users and AI non-users, although AI users saw a smaller range of scores, featuring both a smaller range from minimum to maximum score and a smaller interquartile range. This observation remained relatively stable across six exams.

Of course, that observation as a whole might not mean that AI did not grant an advantage: it could have been the case that AI allowed students who would have underperformed to improve to the level of the class average. If that were the case, however, we would expect that trend to manifest in other ways. For one, we would expect the overall class average to improve if otherwise low-performing students were systematically improving; Figure 1 indicates that is not the case. Second, among those students who use AI on one exam but not the other, we would expect their grade on the exam in which they used AI to be higher; it is not, however.

There may be other reasons why using AI during these exams may grant an advantage. However, there does not appear to be evidence that the most likely outcomes of AI giving students an advantage—either improving their score overall, or preferentially improving scores for otherwise underperforming students—are occurring.

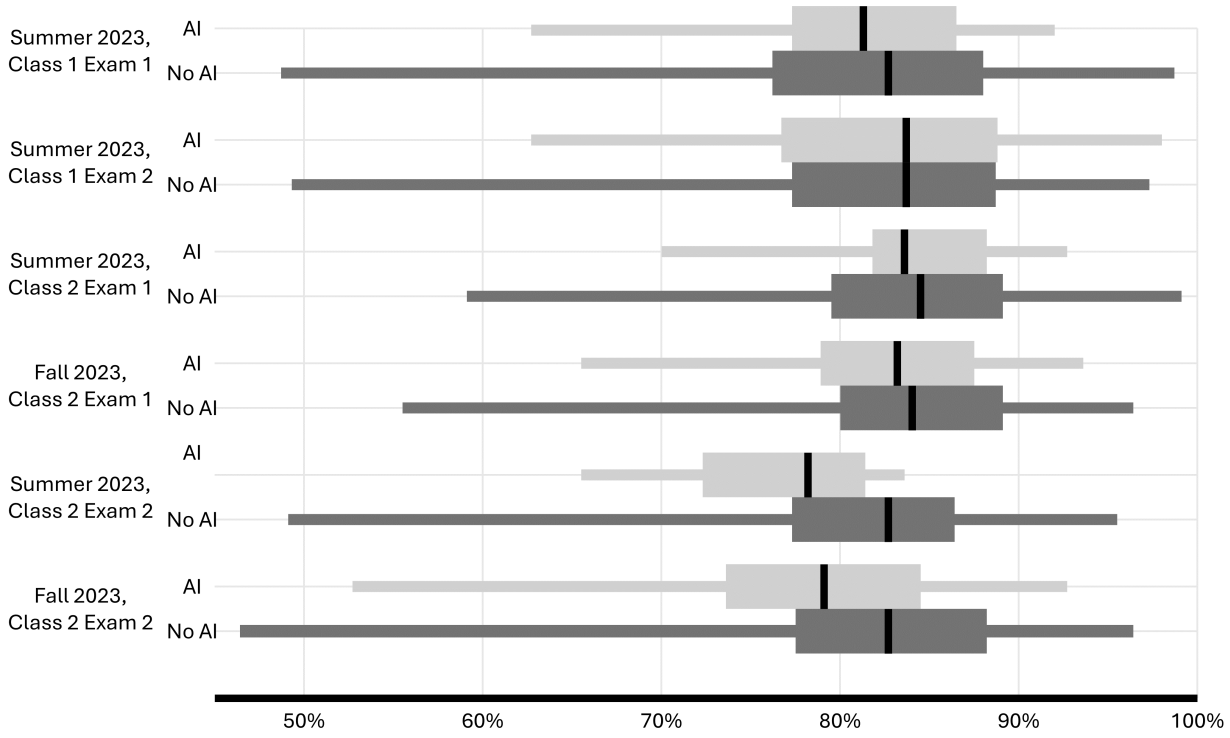
5. ANALYSIS 2: STUDENT STRATEGIES FOR AI USAGE

While the proctoring system recorded the resources students accessed during their exams, the system does not automatically provide an account of *how* students used the resources. For AI resources in particular, while we can systematically measure the extent to which students accessed these resources and correlate that to their overall grades, this analysis alone does not shed light on how students are engaging with these tools. To get an understanding of students' behavior it is crucial to review the screen recordings alongside the list of URLs accessed. This allows for an analysis of the sequence of actions taken by students and the context in which these actions occurred.

Fortunately for this research question (although privacy concerns regarding digital proctoring remain [14]), the proctoring interface also captures recordings of students' exam sessions, including a recording of their screen. In this way, we can investigate how resources are being used: are students copying exam questions into these tools in their entirety, or using them for targeted questioning? Are they getting answers from AI, or asking AI to confirm their earlier guesses? Are they using AI as a first option or a last resort? In this second analysis, we review the exam session recordings themselves to uncover the specific ways that students utilized AI during their exams.

In our review, we noted and analyzed how students used AI during their exams to answer exam questions. This includes

Figure 2: Box plots of the data represented in Table 3. The horizontal axis represents the percentage of questions answered correctly. Black lines represent the median score among students on the given exam in the given category. Plots labeled 'AI' summarize students who used an identified AI tool on their exam; plots labeled 'No AI' summarize students who did not use an identified AI tool on their exam.



recording the circumstances in which AI is applied including the kind of inquiry being addressed and the characteristics of the AI tool utilized such as identifying whether students first refer to lecture slides or other course materials before turning to AI for assistance.

5.1 Methodology

Using the data mined from Summer 2023 for both classes, we identified students who were flagged as using AI. We then selected 22 sessions for manual review: two from Exam 1 in Class 1, five from Exam 2 in Class 1, and fifteen from Exam 2 in Class 2. These selections were driven by multiple constraints: first, permissions to access exam sessions are heavily restricted for privacy protection, and only one researcher with access to exams in Class 1 was available to review, limiting the number of reviewed sessions. Second, exam sessions expire six months after the exam deadline, and this analysis began right as the six-month window ended for Exam 1 in both classes; thus, only Exam 2 could be used in Class 2. Third, fifteen exams for Class 2 comprised all the flagged exams for Exam 2. Of these 22 sessions, the video was corrupted for two sessions and could not be reviewed. Thus, 20 sessions were reviewed, ranging in duration from 90 to 120 minutes. For 17 of these sessions, the AI tool that was flagged was ChatGPT; 6 of these were ChatGPT Plus subscribers. The remaining three sessions accessed Semantic Scholar during their test.

While the proctoring program did record the URLs that were accessed during the exams, it did not record every time the URL was used. Due to this limitation with the proctoring program, we reviewed the entire duration of each screen capture recording to identify each instance where AI was used during the exam. We notated how students used AI and what they used AI for during their exams. After reviewing all assigned recordings, we analyzed the notes taken during each recording for any themes or notable insights. We are particularly interested in identifying common patterns or trends in the approach to using AI tools in answering exam questions.

5.2 Results

After reviewing the notes taken on these sessions, we identified two different ways in which we could categorize AI usage during exams: by how quickly the student resorted to using AI ("Quickness to AI") and by purpose or goal of AI usage ("Function of AI").

Quickness to AI usage refers to the extent to which students treated AI as their first option or if they used it as needed. Although this is a spectrum, we observed that it fairly evenly broken down into three broad categories, which we refer to as "AI First", "AI Second", and "No AI". Function of AI usage refers instead to *how* students were using AI; here we identified two such patterns, which we refer to colloquially as "Trust" and "But Verify".

Interestingly, we did not observe instances of other hypothesized approaches to using AI. We did not see instances of students using AI to check their answers rather than to gen-

erate answers, nor did we see students using AI to help guide them to other sources that might have exact answers. AI was treated by these students as a direct assistant on the exam questions, not an assistant with the content more broadly. It is unclear if that is because this behavior was just absent from the sampled set of sessions, but if these behaviors occur, they appear to be less common than the ones documented here.

In the following sections, we identify students by class and student to track behavior across multiple categories.

5.2.1 Quickness to AI: AI First

Of the 20 sessions we evaluated, eight could be characterized as the student adopting an "AI first" mindset, entering the exam with the intent to use AI from the beginning.

In Class 1, two students used AI, specifically ChatGPT, to attempt all exam questions. Early in the process of taking the exam, these students realized that they could not copy the exam questions and paste them into ChatGPT as right-clicking and copying was disabled by the proctoring software. Due to this limitation, the students adapted and used image-to-text software to convert images of the exam into text that could be pasted into ChatGPT. In order to get ChatGPT to properly answer the exam questions, both the students fed information to ChatGPT.

Student A (from Class 1) first gave ChatGPT the rules of the exam. Then, they gave ChatGPT what appeared to be a copy of lectures notes and transcripts before having ChatGPT answer questions based on the lectures. The exam also tested students on reading material from the course. Initially, the student struggled to locate the reading materials from the course, but when they found the reading material, the student uploaded the reading material to askyourpdf.com and gave the link that the website generated to ChatGPT. ChatGPT did provide answers to the questions based on the readings, but the student did struggle to get ChatGPT to understand some of the questions and papers. The student used the answers ChatGPT provided to answer all the questions, but they did attempt to answer one question by themselves before eventually using the answer ChatGPT provided. This student performed 8% below the class average on the exam.

Student B (also from Class 1) had a similar experience to the first student. They also gave ChatGPT test instructions, but they provided a list of the readings instead of uploading the reading material to ChatGPT. Once the student realized that they could not copy questions, they then tried to print the test, which was unsuccessful, before taking pictures of the test and using image-to-text software. The student provided all the questions to ChatGPT and used the answers it provided, but there were moments when the student disagreed with ChatGPT. The exam was a multiple choice exam where one to four answers could be correct. If ChatGPT only said that one answer was correct, the student would remind ChatGPT that more than one answer was allowed. The student would also question ChatGPT. The assumption is that the student felt like ChatGPT did not provide a correct answer, but that cannot be definitively determined based on the screen recording. Interestingly, ChatGPT attempted to provide answers to questions that the student had not given it. Despite these issues with ChatGPT, the student used the answers provided by ChatGPT to answer the exam questions. This student scored exactly the class average on the exam.

In Class 2, six total students used ChatGPT throughout the entire exam. Three of these students (Students C, D, E) entered exact exam questions into ChatGPT and submitted its exact answers; the other three (Students F, G, and H) used ChatGPT throughout the exam, but separately confirmed its answers as described under the Function of AI: But Verify section below. All six of these students used similar strategies to those used in Class 1 to be able to copy exam questions. One student copy and pasted text from screenshots of exam questions using MacOS Preview before switching to copying the question text from the source code of the exam web page. Students C, D, and E scored 2% below, 13% below, and 3% above the class average, respectively. Students F, G, and H scored 4% above, 4% below, and 7% above the class average, respectively.

5.2.2 Quickness to AI: AI Second

Of the 20 sessions we evaluated, six we instead categorized as "AI Second". These students did not use AI throughout the exams; instead, they did so only when they seemingly encountered a question they felt they needed AI assistance to complete.

In Class 1, only one student—Student I—used ChatGPT to assist them in this way during Exam 1. The student accessed ChatGPT four times during the exam. The student first tried to copy and paste an exam question into ChatGPT, but that method failed since right-clicking and copying was disabled in the proctoring program. They did not type anything into ChatGPT afterwards. The second time, the student asked for a definition from ChatGPT since that is what the exam question was asking. The student attempted to use ChatGPT again, but they never finished typing their question into ChatGPT. For the final time the student accessed ChatGPT, they did ask ChatGPT a question. They typed in a reworded version of the exam question. Overall, this student only used ChatGPT meaningfully two times, and one was to solicit information more than ask it an exam question. This student scored 11% below the class average on the exam.

In Class 2, five students—Students J, K, L, M, and N—used ChatGPT for part of the exam. Student J appeared to intend to use AI from the start but experienced a problem with one AI resource, ChatPDF. The course's eBook was too large to upload to that site, so the student uploaded it to ChatGPT instead. After getting two lengthy and indefinite responses on ChatGPT's interpretation of the course text, this student solely used the course eBook and their own course notes for the remainder of the exam. Students K, L, and M students used ChatGPT as a supplement to other resources, primarily to request a definition of course vocabulary words. Student K used it to define a common English literary term they were unfamiliar with, filling in a knowledge gap unrelated to the course materials. Still, this student and the others in this section generally confirmed their answers from ChatGPT with other sources. Student

M asked ChatGPT one question during the exam but did not return to the tab where it was open and therefore did not read or use its response. Interestingly, these five students scored on average 8% higher than the class average.

5.2.3 *Quickness to AI: No AI*

Among the students who were identified as accessing an AI resource, some were confirmed upon further review to have not actually done so. These false positives are useful both as context for the measurement of AI usage given above as well as for information for others seeking ways to monitor for AI usage.

In Class 1, three students—Students O, P, and Q—visited a site that has an AI component (Semantic Scholar), but did not themselves use that AI to assist them during the exam. These students only went to this web site to access a course reading that was available on the site.

In Class 2, three students—Students R, S, and T—were flagged for AI usage due to having ChatGPT *open*, but it remained in the background. None of those students used it or any other AI-based resources during the exam. It is unclear if the students intended to use it upon entering the exam or if it was simply in the background as part of other activities.

5.2.4 *Function of AI: Trust*

Among the two purposes we identified for students' use of AI, the first we label "Trust" to refer to the fact that these students generally trusted whatever output they received from the AI.

Among the eight students (Students A through H) who used ChatGPT throughout the entire exam session, five—Students A, B, C, D, and E—were identified as predominantly trusting whatever answer was given by ChatGPT. Student J may have intended to operate in this way as well, but technical issues prevented them from doing so. These students generally focused their attention on getting ChatGPT to answer in the structure of the actual exam—some reminded it that each statement could only be true or false, and that multiple statements in a particular question could be true. Students D and E further modified their prompts to reduce the length of ChatGPT's answers. These students started the exam very close to the deadline and wanted ChatGPT to return only "true" or "false" for each question.

5.2.5 *Function of AI: But Verify*

We labeled the second purpose we identified for students' use of AI as "But Verify" to refer to these students' tendency to use AI as the original producer of an answer, but to independently confirm the answer prior to using it.

Three students who used ChatGPT during the entire exam—Students F, G, and H—used the tool more as a supplemental resource. While these students asked ChatGPT for its answers to the vast majority of questions, they also regularly confirmed ChatGPT's answers with other resources, such as the course eBook, lecture videos, and Google. Similarly, Students I, K, L, and M used ChatGPT as a resource to consult, but either separately confirmed the answers they received or asked questions to help *them* select answers in the first place, such as requesting definitions to better understand what a question was asking.

5.3 Discussion

This analysis yields three main takeaways: two on ways of structuring how we categorize students who use AI during exams, and one on the difficulty with automated detection of AI usage.

First, we can categorize students by whether they are using AI as their first option on the exam ("AI First") or their secondary option ("AI Second"). Students in this first category enter the exam seemingly with the *a priori* intention of using AI throughout; students in this second category are familiar enough with AI as a tool to use it as needed, but do not appear to have an advanced intention to rely on it heavily.

Second, we can categorize students by the extent to which they are using AI as a machine for answers or as an aid in generating answers alongside other resources. Some students seemingly operate almost in parallel with Searle's Chinese room [63]: they enter the questions into the tool, and copy answers out of it, but demonstrate little understanding of the content underlying these answers. Other students use AI more as a starting point or assistant: they generally ask it the questions that the test is asking them, but they separately verify or investigate the answer they receive. Although our sample is too small here to generate generalizable conclusions, we nonetheless have the initial observation that students engaging in this latter process appear to outperform the class average, while students engaging in this former process appear to underperform relative to the class average.

Third, around a third of the sessions identified as having evidence of AI collaboration did not actually feature any such usage. These students either used sites that have AI features but did not use those features, or had AI sites in the background but never accessed them directly. This suggests that automated detection of AI usage is not entirely reliable.

6. ANALYSIS 3: USAGE OF OTHER RESOURCES

As a consequence of mining these exams for data pertaining to collaboration with AI, we also derived a dataset of other resources that were accessed during exam sessions. This dataset provides a quick way to look at what other tools students find useful during exams.

6.1 Methodology

As an initial step in mining student exam sessions for evidence of access to AI assistance, the methodology of the previous two analyses first accessed *all* URLs that were accessed and then filtered those to pay attention only to the URLs that indicated AI usage. For this analysis, we reverted back to the dataset prior to this filtering, giving all the URLs that were accessed and maintaining the link to grade data.

In this analysis, we derive some general correlations between resource usage and grace achievement. We also analyze the

types of resources and frequency of usage to establish patterns and recommendations going forward.

The first sub-analysis focuses on comparing resource usage and achievement scores. Two correlation matrices were generated using a Pearson metric. The first was unique URLs vs. grades to determine if any individual resource positively or negatively correlated with exam performance. The second was frequency of resources and exam performance to see if more resource usage led to increased performance on exams.

The second sub-analysis focuses on exploring which distinct resources students generally use during test taking. We utilized general frequency data for unique URLs used by students to find the resources most visited by students. The most accessed resources were then further analyzed to determine how students used these resources.

6.2 Results

Due to some of the course-specific details of the exams in these two classes (for example, exams in Class 1 assess required readings while exams in Class 2 focus only on lecture material), we divide these results by class.

6.2.1 Class 1 Results

For exam 1 in Class 1, the most widely utilized and frequently accessed domains were edstem.org (home of both the official course discussion forums and the course lecture material), google.com, docs.google.com, github.com, the course syllabus web site, and en.wikipedia.org. edstem.org comprised the largest fraction of these at 54.31%, suggesting a high preference for finding answers in official course materials. google.com comprised 24.56% of instances, suggesting more general search for information. Beyond that, the next largest share belonged to docs.google.com at only 2.98%; notably, some official course notes and transcripts are stored in Google Docs, and so this access pattern may indicate either accessing official course notes or accessing students' own resources. The only other sources comprising more than 1% were github.com, the course syllabus, and Wikipedia.

Exam 2 in Class 1 followed a similar pattern, with a handful of additional resources—dl.acm.org and programs.sigchi.org were the major new additions, driven by Exam 2 featuring questions on additional readings. These generally fall similarly into the category of accessing official materials required as part of enrollment in the class, similar to edstem.org and the course syllabus.

For both exam 1 and exam 2 of Class 1, the most widely used domain was edstem.org, indicating that despite the availability of other online resources, students still utilized the class specific content provided on the official course forum as the primary resource for answering exam questions.

Table 4 highlights the overall frequency statistics for resources accessed during exam 1 and exam 2. For both exam 1 and exam 2, Pearson coefficients were derived to determine whether a positive or negative correlation existed between student scores and frequency of resource usage. The resulting Pearson coefficient for exam 1 with respect to the correlation between score and frequency of resource usage was 0.08; the resulting Pearson coefficient for exam 2 with respect to the correlation between score and frequency of resource usage was 0.11. Each of these resultant Pearson coefficients fall within the interval [0, 0.2], which indicates a very weak positive correlation between student score and frequency of resource usage during exams.

Table 4: Average number of resources accessed per student during each exam; for example, the average student on Exam 1 accessed 31.86 different URLs during their exam (excluding the exam URL itself).

		Average	Stdev
Summer 2023	Exam 1	31.86	34.24
	Exam 2	37.47	40.58

With regard to exam 1, the Pearson coefficients between score and each of the URL domains whose usage frequency >1% is as follows: 0.08 for edstem.org, -0.01 for google.com, 0.07 for docs.google.com, 0.02 for github.com, 0.11 for the course syllabus, and 0.09 for en.wikipedia.org.

With regard to exam 2, the Pearson coefficients between score and each of the URL domains whose usage frequency >1% is as follows: 0.09 for edstem.org (the official course forum and lecture material), -0.05 for google.com, 0.11 for docs.google.com, 0.09 for dl.acm.org (where some required readings are hosted), 0.06 for programs.sigchi.org (home of other required readings), -0.05 for github.com, and 0.07 for the course syllabus.

With respect to the correlation between score and the usage frequency of specific URLs in general, the resulting Pearson correlation coefficients between score and each unique URL used across both exam 1 and exam 2 each fell within the interval [-0.2, 0.2], indicating the presence of only very weak positive and very weak negative correlation strengths. It therefore does not appear to be the case that more frequent resource access leads to higher or lower grades, either as a whole or using any specific resource.

6.2.2 Class 2 Results

Results for Class 2 mirrored those for Class 1: the most frequently-used resources were again the official course forum and lectures at edstem.org, as well as google.com as a whole. No notable correlations between frequency of accessing any specific resource, or resources as a whole, were observed. Interestingly however, Class 2 saw a notably greater fraction of students using translate.google.com; this suggests that the narrow time constraints of the test may have an outsized impact on ESL students who need to rely on translators to interpret questions. Here still, however, the correlation between frequency of resource access and exam performance was week. Fall 2023 again saw similar ratios, but chat.openai.com emerged as an additional high-use resource.

Table 5 highlights the overall frequency statistics for resource access. Two results stand out from the data for class 2. Resource usage on exam 2 is consistently less than exam 1 for all cohorts, potentially due to optimizing resource usage as students become more familiar with test formatting and expectations. The other interesting result stems from

students' usage of Google translate, highlighting a subgroup of students that are utilizing translation software to better understand questions as ESL students.

Table 5: Class 2 Resource Usage Frequency Statistics

		Average	Stdev
Summer 2023	Exam 1	53.4	64.8
	Exam 2	40.8	57
Fall 2023: Section 1	Exam 1	36.2	42.1
	Exam 2	19.3	32.8
Fall 2023: Section 2	Exam 1	52.1	254.8
	Exam 2	22.9	34.6

7. DISCUSSION

Since the release of ChatGPT and similar tools, teachers have been scrambling to adjust their assessments to the knowledge that students now have access to widespread and sophisticated assistance beyond what they have had in the past. To engage in this adjustment properly, however, it is important to understand how students actually *are* using these tools. It is possible, of course, that students are using these tools to circumvent the learning goals of their assessments and to appear as if they understand more than they do. It is also possible, though, that they are using these tools are learning resources or assessment assistants, allowing them to bring out more of their knowledge or develop that knowledge faster. We see some evidence for this idea in Analysis 3: while automatic language translation has been around longer than tools like ChatGPT, it too is an example of an AI tool students may use. Prohibiting use of all AI tools without understanding how they are being used risks missing out on some key benefits they can provide.

This study has attempted to put down some early lines about how students use AI assistance on assignments where they are permitted to do so, but where we also can carefully track how those tools are being used and what impact they are having on student scores. What we have seen here is evidence that the impact of such tools may not be particularly high: Analysis 1 finds that even when such tools were explicitly permitted, a minority—a large one, granted—of students chose to use them. Those students who chose to use these tools did not outperform students who did not, nor did those students who chose to use them some of the time do any better when they used these tools than when they themselves did not.

Analysis 2 gives a more thorough investigation of how students are using these tools, and finds an array of different patterns. Some students enter exams seemingly intent to use AI on the entire assessment, while others default to using AI when they encounter specific questions or obstacles they think AI can help them answer or overcome. Some students trust the AI's answers entirely, while others use it as a piece of a broader problem-solving exercise. Encouragingly, early evidence appears to suggest that these students who use it as a complement to their own knowledge derive a benefit, while those students who use it as a replacement of their own knowledge underperform relative to the rest of the class. Significant additional research is necessary to explore the extent to which this trend holds over larger numbers of exam sessions, as well as other types of assignments.

7.1 Limitations & Future Work

Of course, significantly more work is necessary. One of the exciting and challenging elements of researching this area is that it is constantly changing: we see student use of AI in one class in this study skyrocket between Summer and Fall semesters with no discernible cause, especially given that use in another class during that Summer semester was already high. We have seen increasing evidence of a phenomenon in at-scale education where tactics for certain assignments become commoditized. In more recent semesters, we have witnessed students building custom GPTs to act as exam assistants, a behavior we generally consider acceptable—even desirable—when the process of building one's own GPT provides valuable learning outcomes on its own. With this, however, we expect to see other students borrow their classmates' prebuilt agents, deriving some of the benefits with none of the learning outcomes. When this becomes common, we expect to have to institute a rule where students are only allowed to use AI agents that they themselves created—a policy whose necessity would have been unheard of only a few years ago.

Aside from the fact that this domain is changing rapidly, this study has also only examined these issues in a narrow context: on multiple-choice exams in graduate-level computer science classes in an online Master's program. This student body does not generalize: they are more familiar with AI as a whole and thus likely more likely to be comfortable using these tools, but they also have more intrinsic motivations to learn [13] and thus may be less likely to find ways to circumvent the goals of different assessments. More work is needed to assess how these trends generalize to other audiences, domains, and levels. Toward this end, this study also gives a framework for conducting these evaluations: we argue it is useful to offer assessments that permit AI assistance, but to deliver them through mechanisms that allow for careful monitoring of such assistance to understand how it is being used and to adjust accordingly.

8. ACKNOWLEDGMENTS

Data used in this analysis was gathered and analyzed in accordance with institute IRB protocol H15249.

9. REFERENCES

[1] I. Adeshola and A. P. Adepoju. The opportunities and challenges of ChatGPT in education. *Interactive Learning Environments*, pages 1–14, 2023.

[2] K. L. Adkins and D. A. Joyner. Scaling anti-plagiarism efforts to meet the needs of large online computer science classes: Challenges, solutions, and recommendations. *Journal of Computer Assisted Learning*, 38(6):1603–1619, 2022.

[3] F. Ahmed, K. Shubeck, and X. Hu. ChatGPT in the generalized intelligent framework for tutoring. In *Proceedings of the 11th Annual Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym11)*, page 109. US Army Combat Capabilities Development Command–Soldier Center, 2023.

[4] M. A. AlAfnan, S. Dishari, M. Jovic, and K. Lomidze. ChatGPT as an educational tool: Opportunities, challenges, and recommendations for communication,

business writing, and composition courses. *Journal of Artificial Intelligence and Technology*, 3(2):60–68, 2023.

[5] S. Arnò, A. Galassi, M. Tommasi, A. Saggino, and P. Vittorini. State-of-the-art of commercial proctoring systems and their use in academic online exams. *International Journal of Distance Education Technologies (IJDET)*, 19(2):55–76, 2021.

[6] S. Aurelia, R. Thanuja, S. Chowdhury, and Y.-C. Hu. AI-based online proctoring: a review of the state-of-the-art techniques and open challenges. *Multimedia Tools and Applications*, pages 1–23, 2023.

[7] N. Brouwer, A. Heck, and G. Smit. Proctoring to improve teaching practice. *MSOR Connections*, 15(2), 2016.

[8] S. Coghlan, T. Miller, and J. Paterson. Good proctor or "big brother"? Ethics of online exam supervision technologies. *Philosophy & Technology*, 34(4):1581–1606, 2021.

[9] D. R. Cotton, P. A. Cotton, and J. R. Shipway. Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, pages 1–12, 2023.

[10] J. A. Crowder and J. N. Carbone. Collaborative shared awareness: human-AI collaboration. In *Proceedings of the International Conference on Information and Knowledge Engineering (IKE). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp)*, volume 1, 2014.

[11] M. C. DiBartolo and C. M. Walsh. Desperate times call for desperate measures: Where are we in addressing academic dishonesty? *Journal of Nursing Education*, 49(10):543–544, 2010.

[12] B. DiSalvo, D. Bandaru, Q. Wang, H. Li, and T. Plötz. Reading the room: Automated, momentary assessment of student engagement in the classroom: Are we there yet? *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(3):1–26, 2022.

[13] A. Duncan, B. Eicher, and D. A. Joyner. Enrollment motivations in an online graduate CS program: Trends & gender-and age-based differences. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, pages 1241–1247, 2020.

[14] A. Duncan and D. Joyner. On the necessity (or lack thereof) of digital proctoring: Drawbacks, perceptions, and alternatives. *Journal of Computer Assisted Learning*, 38(5):1482–1496, 2022.

[15] B. L. Eicher and D. A. Joyner. Making the grade: Assessments at scale in a large online graduate program. In *2023 IEEE Learning with MOOCS (LWMOOCS)*, pages 1–6. IEEE, 2023.

[16] S. Elbanna and L. Armstrong. Exploring the integration of ChatGPT in education: adapting for the future. *Management & Sustainability: An Arab Review*, 3(1):16–29, 2024.

[17] A. Fügener, J. Grahl, A. Gupta, and W. Ketter. Cognitive challenges in human–artificial intelligence collaboration: Investigating the path toward productive delegation. *Information Systems Research*, 33(2):678–696, 2022.

[18] N. Gaumann and M. Veale. AI providers as criminal essay mills? Large language models meet contract cheating law. 2023.

[19] A. Giannopoulou, R. Ducato, C. Angiolini, and G. Schneider. From data subjects to data suspects: Challenging e-proctoring sytems as a university practice. *J. Intell. Prop. Info. Tech. & Elec. Com. L.*, 14:278, 2023.

[20] A. Goel and D. Joyner. An experiment in teaching artificial intelligence online. *International Journal for the Scholarship of Technology-Enhanced Learning (1)*, 1, 2016.

[21] A. K. Goel and L. Polepeddi. Jill watson: A virtual teaching assistant for online education. In *Learning engineering for online education*, pages 120–143. Routledge, 2018.

[22] J. Goodman, J. Melkers, and A. Pallais. An elite grad-school degree goes online. *Education Next*, 18(3):66–72, 2018.

[23] J. Goodman, J. Melkers, and A. Pallais. Can online delivery increase access to education? *Journal of Labor Economics*, 37(1):1–34, 2019.

[24] S. Grassini. Shaping the future of education: exploring the potential and consequences of AI and ChatGPT in educational settings. *Education Sciences*, 13(7):692, 2023.

[25] R. Graziano, D. Benton, S. Wahal, Q. Xue, P. T. Miller, N. Larsen, D. Vacanti, P. Miller, K. C. Mahajan, D. Srikanth, et al. Jack watson: Addressing contract cheating at scale in online computer science education. In *Proceedings of the Sixth (2019) ACM Conference on Learning@ Scale*, pages 1–4, 2019.

[26] T. Greitemeyer and A. Kastenmüller. HEXACO, the Dark Triad, and ChatGPT: Who is willing to commit academic cheating? *Heliyon*, 9(9), 2023.

[27] C. Gunn. Illinois virtual campus: Focusing on student support. *Technology Source*, 2000.

[28] D. Harwell. Cheating-detection companies made millions during the pandemic. now students are fighting back. In *Ethics of Data and Analytics*, pages 410–417. Auerbach Publications, 2022.

[29] S. Herbold, A. Hautli-Janisz, U. Heuer, Z. Kikteva, and A. Trautsch. A large-scale comparison of human-written versus ChatGPT-generated essays. *Scientific Reports*, 13(1):18617, 2023.

[30] E. Hutchins. Distributed cognition. *International Encyclopedia of the Social and Behavioral Sciences. Elsevier Science*, 138:1–10, 2000.

[31] D. Joyner. Squeezing the limeade: policies and workflows for scalable online degrees. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*, pages 1–10, 2018.

[32] D. A. Joyner. Scaling expert feedback: Two case studies. In *Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale*, pages 71–80, 2017.

[33] D. A. Joyner. Meet me in the middle: Retention in a" mooc-based" degree program. In *Proceedings of the Ninth ACM Conference on Learning@ Scale*, pages 82–92, 2022.

[34] D. A. Joyner. ChatGPT in education: Partner or pariah? *XRDS: Crossroads, The ACM Magazine for*

Students, 29(3):48–51, 2023.

[35] D. A. Joyner and C. Isbell. Master's at scale: Five years in a scalable online graduate degree. In *Proceedings of the Sixth (2019) ACM Conference on Learning@ Scale*, pages 1–10, 2019.

[36] D. A. Joyner, C. Isbell, T. Starner, and A. Goel. Five years of graduate cs education online and at scale. In *Proceedings of the ACM conference on global computing education*, pages 16–22, 2019.

[37] S. Kolowich. Behind the webcam's watchful eye, online proctoring takes hold. *Chronicle of Higher Education*, 2013.

[38] Q. Kreth, M. E. Spirou, S. Budenstein, and J. Melkers. How prior experience and self-efficacy shape graduate student perceptions of an online learning environment in computing. *Computer Science Education*, 29(4):357–381, 2019.

[39] T. Lancaster. Artificial intelligence, text generation tools and ChatGPT–does digital watermarking offer a solution? *International Journal for Educational Integrity*, 19(1):10, 2023.

[40] K. Lee and M. Fanguy. Online exam proctoring technologies: Educational innovation or deterioration? *British Journal of Educational Technology*, 53(3):475–490, 2022.

[41] M. Lee and M. W. Ashton. The calculator of writing. *Australasian Journal of Plastic Surgery*, 6(1):1–4, 2023.

[42] X. Li, K.-m. Chang, Y. Yuan, and A. Hauptmann. Massive open online proctor: Protecting the credibility of moocs certificates. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 1129–1137, 2015.

[43] K. Linden and P. Gonzalez. Zoom invigilated exams: A protocol for rapid adoption to remote examinations. *British Journal of Educational Technology*, 52(4):1323–1337, 2021.

[44] C. K. Lo. What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences*, 13(4):410, 2023.

[45] C. Logan. Toward abolishing online proctoring: Counter-narratives, deep change, and pedagogies of educational dignity. *Journal of Interactive Technology and Pedagogy*, 20, 2021.

[46] S. Manoharan, U. Speidel, A. E. Ward, and X. Ye. Contract cheating–dead or reborn? In *2023 32nd Annual Conference of the European Association for Education in Electrical and Information Engineering (EAEEIE)*, pages 1–5. IEEE, 2023.

[47] M. Montenegro-Rueda, J. Fernández-Cerero, J. M. Fernández-Batanero, and E. López-Meneses. Impact of the implementation of ChatGPT in education: A systematic review. *Computers*, 12(8):153, 2023.

[48] F. Mosaiyebzadeh, S. Pouriyeh, R. Parizi, N. Dehbozorgi, M. Dorodchi, and D. Macêdo Batista. Exploring the role of ChatGPT in education: Applications and challenges. In *Proceedings of the 24th Annual Conference on Information Technology Education*, pages 84–89, 2023.

[49] N. Naikar, A. Brady, G. Moy, and H.-W. Kwok. Designing human-AI systems for complex settings: ideas from distributed, joint, and self-organising perspectives of sociotechnical systems and cognitive work analysis. *Ergonomics*, 66(11):1669–1694, 2023.

[50] A. Nigam, R. Pasricha, T. Singh, and P. Churi. A systematic review on AI-based proctoring systems: Past, present and future. *Education and Information Technologies*, 26(5):6421–6445, 2021.

[51] J. A. Oravec. Artificial intelligence implications for academic cheating: Expanding the dimensions of responsible human-AI collaboration with ChatGPT. *Journal of Interactive Learning Research*, 34(2):213–237, 2023.

[52] D. S. Park, R. W. Schmidt, C. Akiri, S. Kwak, and D. A. Joyner. Affordable degrees at scale: New phenomenon or new hype? In *Proceedings of the Seventh ACM Conference on Learning@ Scale*, pages 25–35, 2020.

[53] M. S. Parsa and L. Golab. Academic integrity during the COVID-19 pandemic: a social media mining study. In *Proc. Fourteenth International Conference on Educational Data Mining (EDM 2021)*, number 73, 2021.

[54] S. Patael, J. Shamir, T. Soffer, E. Livne, H. Fogel-Grinvald, and L. Kishon-Rabin. Remote proctoring: Lessons learned from the COVID-19 pandemic effect on the large scale on-line assessment at tel aviv university. *Journal of Computer Assisted Learning*, 38(6):1554–1573, 2022.

[55] T. Phung, J. Cambronero, S. Gulwani, T. Kohn, R. Majumdar, A. Singla, and G. Soares. Generating high-precision feedback for programming syntax errors using large language models. *arXiv preprint arXiv:2302.04662*, 2023.

[56] M. Pradana, H. P. Elisa, and S. Syarifuddin. Discussing ChatGPT in education: A literature review and bibliometric analysis. *Cogent Education*, 10(2):2243134, 2023.

[57] J. Qi, H. Tang, and Z. Zhu. Exploring an affective and responsive virtual environment to improve remote learning. In *Virtual Worlds*, volume 2, pages 53–74. MDPI, 2023.

[58] M. M. Rahman and Y. Watanobe. ChatGPT for education and research: Opportunities, threats, and strategies. *Applied Sciences*, 13(9):5783, 2023.

[59] J. Rajala, J. Hukkanen, M. Hartikainen, and P. Niemelä. "call me kiran"–ChatGPT as a tutoring chatbot in a computer science course. In *Proceedings of the 26th International Academic Mindtrek Conference*, pages 83–94, 2023.

[60] A. B. Redding. Fighting back against achievement culture: Cheating as an act of rebellion in a high-pressure secondary school. *Ethics & Behavior*, 27(2):155–172, 2017.

[61] S. Rodchua, G. Yaiadom-boakye, and R. Woolsey. Student verification system for online assessments: Bolstering quality and integrity of distance learning. *Journal of Industrial Technology*, 27(3), 2011.

[62] Z. H. Sain and M. T. Hebebci. ChatGPT and beyond: The rise of AI assistants and chatbots in higher education. *Technology and Education*, page 1, 2023.

[63] J. R. Searle. The Chinese room revisited. *Behavioral and brain sciences*, 5(2):345–348, 1982.

[64] N. Selwyn, C. O'Neill, G. Smith, M. Andrejevic, and

X. Gu. A necessary evil? The rise of online exam proctoring in australian universities. *Media International Australia*, 186(1):149–164, 2023.

[65] S. Silverman, A. Caines, C. Casey, B. Garcia de Hurtado, J. Riviere, A. Sintjago, and C. Vecchiola. What happens when you close the door on remote proctoring? Moving toward authentic assessments with a people-centered approach. *To Improve the Academy: A Journal of Educational Development*, 39(3), 2021.

[66] J. L. Steele. To GPT or not GPT? Empowering our students to learn with AI. *Computers and Education: Artificial Intelligence*, 5:100160, 2023.

[67] S. Swauger. Our bodies encoded: Algorithmic test proctoring in higher education. *Critical digital pedagogy*, 2020.

[68] C. E. Tatel, S. F. Lyndgaard, R. Kanfer, and J. E. Melkers. Learning while working: Course enrollment behaviour as a macro-level indicator of learning management among adult learners. *Journal of Learning Analytics*, 9(3):104–124, 2022.

[69] M. Virvou and G. A. Tsihrintzis. Is ChatGPT beneficial to education? A holistic evaluation framework based on intelligent tutoring systems. In *2023 14th International Conference on Information, Intelligence, Systems & Applications (IISA)*, pages 1–8. IEEE, 2023.

[70] D. Wang, D. Shan, Y. Zheng, K. Guo, G. Chen, and Y. Lu. Can ChatGPT detect student talk moves in classroom discourse? A preliminary comparison with Bert. In *Proceedings of the 16th International Conference on Educational Data Mining*, pages 515–519. International Educational Data Mining Society, 2023.

[71] Q. Wang, S. Jing, D. Joyner, L. Wilcox, H. Li, T. Plötz, and B. Disalvo. Sensing affect to empower students: Learner perspectives on affect-sensitive technology in large educational contexts. In *Proceedings of the Seventh ACM Conference on Learning@ Scale*, pages 63–76, 2020.

[72] Q. Wang, S. E. Walsh, M. Si, J. O. Kephart, J. D. Weisz, and A. K. Goel. Theory of mind in human-ai interaction. *interactions*, 28:33, 2024.