Bob or Bot: Exploring ChatGPT's answers to University Computer Science Assessment

Bob or Bot: Exploring ChatGPT's answers to University Computer Science Assessment

Mike Richards, Kevin Waugh, Mark Slaymaker, Marian Petre, Daniel Gooch

School of Computing and Communications, The Open University, UK. Contact mike.richards@open.ac.uk

Cheating has been a problem long standing issue in University assessments. However, the rise of ChatGPT and other free-to-use generative AI tools have democratised cheating. Students can run any assessment questions through the tool, and generate a superficially compelling solution, which may or may not be accurate. We ran a blinded "quality assurance" marking exercise, providing ChatGPT-generated "synthetic" scripts alongside student scripts to volunteer markers. 4 end-of-module assessments from across a University CS curriculum were anonymously marked. A total of 90 scripts were marked, and barring two outliers, every undergraduate script received *at least* a passing grade. We also present the results of running our sample scripts through diverse quality assurance software, and the results of interviewing the markers. As such, we contribute a baseline understanding of how the public release of generative AI is potentially going to significantly impact quality assurance processes as our analysis demonstrates that, in most cases, across a range of question formats, topics, and study levels, ChatGPT is at least capable of producing adequate solutions.

CCS CONCEPTS • **Applied computing~Education**~Distance learning• **Applied computing~Education** • **Computing methodologies~Artificial intelligence**

Additional Keywords and Phrases: ChatGPT, Generative AI, Cheating, Quality Assurance, University Assessment,

1 INTRODUCTION

In late November 2022, OpenAI opened public access to its ChatGPT program. ChatGPT showed unprecedented capabilities in generating textual responses to prompts and caused a sensation – not only in academia and the fields of artificial intelligence and natural language understanding, but also amongst the wider public. Within hours of the announcement, newspapers, magazines and online sites were posting excited articles detailing their experiments with ChatGPT.

The enormous training set underpinning GPT-3 [21] allows ChatGPT to generate relatively large volumes of fluent text, that – at least superficially – resemble that written by human authors. Such is the breadth of the data that a wide range of potential problem domains have been proposed for ChatGPT – including summarising complex documents, chatbots for customer enquiries, intelligent tutors and in the role of a journalistic writer. Unlike conventional web search engines

ChatGPT is a closed system. It does not acquire new data by a process of web crawling and is therefore unable to provide information about recent events[1]. Therefore, it has limited use as a tool for responding to current affairs.

LLMs lack any “understanding” of the meaning of their training data, that is they do not maintain any internal representation of the world. For this reason, they have been described as “*completely mindless”* [23] and *“stochastic parrots”* [2] capable of producing large volumes of text that may be misleading, contain biases derived from inherent flaws in their training sets, or entirely wrong. However, such is their apparent fluency, that users and readers alike can be seduced into believing that these responses are the result of genuine intelligence founded on facts.

The release of ChatGPT has not gone unnoticed in the academic community. In addition to discussion within the academy, the wider media has also published many articles about the impact of ChatGPT on education, particularly assessment [44]. ChatGPT provides a new approach for students to cheat, by inputting assessment questions, and receiving ‘synthetic solutions’ which can be passed-off as a student’s work.

Academic integrity underpins institutional reputations. As Dick *et al.* point out, awarding academic qualifications to dishonest students has serious consequences – not least to the reputation of the institution itself, but also to wider society: it damages the value placed on all academic qualifications; it damages the reputation of associated professions; and perhaps most seriously, it could endanger society as incompetent graduates enter employment and produce substandard or dangerous work [15].

To demonstrate empirically the risk ChatGPT poses to academic integrity, we ran a “quality assurance” marking exercise using assessments from a range of Computer Science modules. Our results highlight that ChatGPT produces average answers that would pass most of the assessments used, and that standard means of misconduct detection are not sufficient to identify synthetic solutions. We conclude by arguing our analysis demonstrates the need to better explore what assessment types for which topic areas are most resistant to generative AI, to ensure academic quality assurance, while continuing to support students’ learning.

2 LITERATURE REVIEW

All the authors are based at the Open University (OU) in the UK. As a distance education provider, the OU has a particular approach to teaching which is necessary to outline, before considering the broad literature on academic misconduct.

2.1 How the Open University teaches

The Open University (OU) is a large distance education university based in the United Kingdom. With more than 200,000 active students, it is the largest university in the UK awarding undergraduate and postgraduate degrees, as well as non-degree qualifications such as diplomas and certificates or continuing education units. Except for PhD students, the majority of OU students study part time at a distance using especially designed study materials developed by academic module

[1] At the time of writing (March 2023), the ‘horizon’ for ChatGPT is September 2021. Queries for events after this date receive a response such as: “I'm sorry, but as an artificial intelligence language model, I do not have access to real-time information or external data sources. Additionally, I do not have the capability to observe events or actions that occur in the physical world.”

teams. Module teams are formed of central academics, who, in addition to developing the study materials, also write the assessment material.

OU modules are designed to be delivered at very large scale with presentations of individual modules often exceeding 1,000 students. This is made possible through the role of Associate Lecturers (ALs) who each lead one or more tutor groups each containing up to 25 students. ALs have several responsibilities, including:

- Marking assessment following guidance from the module team;
- Providing additional teaching through tutorials;
- Acting as a point of contact for students both for study and pastoral purposes.

Award-bearing qualifications from the OU are assessed both *during* individual modules through one or more Tutor Marked Assignments (TMA, broadly equivalent to coursework) and at the *end* of modules by an End of Module Assignment (EMA, similar to coursework but completed as the end point assessment) or exam.

The vast majority of student TMAs and EMAs are submitted electronically to the University's assessment portal, where they can be retrieved for marking by ALs. As a response to the Covid-19 pandemic, the OU moved from traditional invigilated face-to-face examinations to online open book examinations. Markers are typically allocated 75-100 EMA/exam scripts, determined by the size of the module cohort.

A student's EMA or exam script is marked by an AL teaching on the module who has not been providing tutoring to that student. Other quality control measures include using automated plagiarism checks (TurnItIn and CopyCatch). Individual ALs can flag assessment solutions for review by module teams if they suspect content has been plagiarised, have other academic conduct concerns, or they do not feel qualified to mark a solution. All assignment marking undergoes a monitoring process to guarantee quality. Depending on the type of module, monitoring might be restricted to remarking or double marking a random sample of solutions and noting divergences; or having every script marked by multiple markers.

2.2 Academic misconduct

Cheating has been a threat to the integrity of education for as long as there have been examinations. In our own discipline of computer science, the conversation around cheating extends back several decades [16].

As Dick *et al.* point out; a simple definition of cheating is somewhat elusive; however, they proposed the following:

"A behavior may be defined as cheating if one of the two following questions can be answered in the positive.

- *"Does the behavior violate the rules that have been set for the assessment task?*
- *"Does the behavior violate the accepted standard of student behavior at the institution?"* [15]

Furthermore, Dick *et al.* emphasise that institutions must make students aware of what is – and what is not – acceptable behaviour, using systems such as student codes of conduct.

It is extremely difficult to determine the scale of cheating in academia due to the natural fear that students who admit to the practice may be penalised. Based on a randomised, privacy-preserving response, a study by Brunelle & Hott [9] surveyed students on an algorithms course (847 students) of whom 41.68% admitted cheating; and a CS theory course (181 students) in which half of the cohort was asked about cheating in coding (of whom 29.27% cheated) and half was asked about cheating in written assignments (31.31% cheated).

Over a range of practices, the main influences on cheating in significant, large studies were found to be time pressure and concerns about failing, while the main countering influences were students' desire to learn and to know what they have learnt [33,34]. Overemphasis on high grades may encourage students to engage in dishonest behaviours; Dyer *et al.* describe a need to develop greater academic integrity: *"It [academia] needs to address the perception of higher education today as transactional in nature and of the need to get good grades as more important than the acquisition of knowledge"* [17].

Some academic institutions have attempted novel assessment strategies to defeat – or at least deter – cheating; examples include generating unique exams for each student [38] or generating unique multiple-choice quizzes by selecting questions randomly from a large pool [12]. These solutions prevent dissemination of answers between students sharing a common piece of assessment but are much less effective when LLMs can produce appropriate bespoke solutions to individual students within a few seconds.

In the following subsections we consider how cheating in particular assessment formats have been dealt with by the STEM education community.

2.3 Assessment formats

2.3.1 Essay mills

Essay mills are services where students commission assessment solutions from individual writers engaged on an *ad hoc* basis [34]. A study by Newton (2018) found that 15.7% of students surveyed admitted to paying someone else to write an assignment.

Following the introduction of UK legislation criminalising the advertising or operation of essay mills [13,31], the Quality Assurance Agency for Higher Education (QAA) published updated guidance on essay mills [34]. The key recommendations are unsurprising: providing clear information and support for students; ensuring sufficient training and support for staff to identify material produced by mills; reducing opportunities for use; a requirement for strong detection measures and the development of clear and accessible institutional regulations and policies.

Research into the role of essay mills has produced mixed results. Some report poor quality submissions. For example, Sutherland-Smith and Kevin Dullaghan submitted 58 orders to 18 distinct essay mill providers. 30% of the outputs lacked specified components or were of low quality (missing sections or late), with 52% of the purchased work failing to reach the required passing grade [45].

Conversely, Lines reported on the submission of undergraduate and Masters essays (200 words apiece) from 13 sites [28]. Whilst two of the undergraduate essays failed, ten passed, including two receiving credit and one a high distinction. For the Masters-level course, six failed and seven passed, with two submissions receiving credit, one a distinction and one a high distinction. The researchers note that, roughly, the price paid for work corresponded to the mark received.

2.3.2 Examinations

The exam format has survived because it is efficient to administer, cheap to mark, and the controlled examination environment creates challenges for those choosing to cheat or plagiarise work. For distance teaching institutions, (including the OU), face-to-face examinations serve a crucial secondary purpose of authenticating student identities by requiring exam takers to provide official photographic identity (such as a passport, driving licence or identity card). Despite this authentication purpose, not all areas of computing are well-suited to the traditional exam format; for example, an understanding of programming is best examined through practical activities on a computer [11].

A disadvantage of the conventional exam format, especially for distance educators, is the considerable cost imposed on students to travel to the examination centres, and the cost to institutions in hiring suitable venues. Remote examinations, where a student completes the assessment at a location other than a centralised examination centre, eliminates these costs entirely, and can offer flexibility to students in when and how they sit their assessment. Remote online proctored examinations became a necessary feature of many institutions' responses to the pressure of the Covid-19 pandemic.

Security for remote exams can be increased through the process of "proctoring"; defined by [25] as *"the use of virtual tools for monitoring student activities during assessment activity"*. In a proctored exam, students may be required to use a webcam to "scan" the location where they are taking the exam to demonstrate that no unauthorised study materials, devices or other people are present. Perhaps unsurprisingly, unproctored remote examinations have been linked to increased student scores. Schultz *et al.* point out that unproctored remote examinations are effectively open book examinations, since the students can bring additional materials into their personal exam environment whether or not that is permitted by the institution [40].

Proctoring has been strongly contested as a gross, and illegal, violation of privacy [5] as well as discriminatory, by assuming students have access to a reasonably private place where they can work on an exam without disturbance or violating the rights of others. As Swauger explains, there are concerns that proctoring technologies employing biometric techniques (*e.g.*, facial recognition) disproportionately disadvantage minority students [46].

Additionally, trials have demonstrated that AI proctoring software is a poor detector of cheating behaviour. Bergmans *et al*. [4] present a controlled experiment using the AI monitoring software Proctorio, which can flag "suspicious behaviour" and alert a human invigilator to intervene. The experiment recruited 30 computer science students, 5 of whom were asked to behave nervously but complete the test honestly, and a further 6 students who were asked to cheat. The system was ineffective at identifying cheating students, with nervous students also being flagged.

Traditional formats of summative assessments such as written final exams are known to cause anxiety and continued frustration, which may influence students' self-efficacy beliefs, whilst open-book exams can be of great benefit to students, reducing anxiety and stress [36]. In such cases, highlighting institutional plagiarism policies and highlighting academic good practice has been associated with reduced plagiarism in CS courses [29,30].

2.3.3 Oral examinations (vivas)

Replacing written assignments with oral examinations has been demonstrated in CS, with mixed feedback from students [27] – although it should be pointed out that all forms of assessment attract a diverse range of opinions. Motivated by reducing plagiarism in a system analysis and design course, Dick studied student interviews as a means of assessment [14] and found that performing interviews at two points during a semester-long team project eliminated the student practice of copying assignments and disguising plagiarism by making minor changes. Further positive findings were that students received immediate feedback and had opportunities to practice and develop their communication skills. On the negative side, many students found the interview process stressful and expressed a strong dislike for this form of assessment.

Interested in improving the effectiveness of providing student feedback, East and Schafer studied the implementation of in-person grading using individual meetings between instructor and student to discuss and evaluate the student's work [18]. Students who participated in these personal grading sessions expressed their preference for this assessment method, not only finding them of use, but also suitable for assessment purposes.

Ohmann discovered that the students and instructors who participated in a final oral exam in a foundational CS course had positive reactions about their experience, with students demonstrating a deeper level of engagement with the material than previously noted with written assessment [32]. Regarding barriers to the implementation of oral exams, Ohmann cited difficulties with scaling the exam to a larger class, especially when single instructors do not have sufficient support from tutors or teaching assistants. Ohmann highlighted a significant problem with students from distributing exam questions to peers during the five-day period required to complete every individual session.

Motivated by the unprecedented imposition of lockdowns during the COVID-19 pandemic, Sabin *et al.* presented work related to a forced transition from written in-class exams to oral assessments conducted at a distance [39]. Most students reported that they felt the oral examination was either comparable to the written assessment or easier than prior assessment. Students taking the oral exam reported they spent over 50% longer preparing for the examination compared to those taking written exams. The experimenters suggested that oral examinations drive deeper student engagement with materials because students felt they required a greater understanding to have a conversation with an informed examiner.

2.4 The impact of LLMs on assessment

Writing in the immediate aftermath of ChatGPT's dazzling introduction, it is hard to determine the future role of LLMs in education; but, come what may, this technology is here to stay. It is incumbent on educators to teach our students what LLMs are, how they work, and how to make the best use of them [41]. LLMs potentially offers many educational benefits, including: the potential to act as 'informed' study companions – especially useful for students working alone and in distance education contexts; simplifying developing personalised assessment; supporting creative writing; and engaging in the iterative development of a program.

However, our focus is on the risks generative AI poses to academic conduct. Software such as ChatGPT is extremely convenient, easy to use, and low (or zero) cost – effectively democratising the ability for students to cheat.

2.4.1 The performance of LLMs in academic assessment

The relatively recent public availability of LLMs such as ChatGPT means that there is little in the published peer-reviewed literature that outlines the capabilities, shortcomings, benefits, and risks of the technology for educational purposes based on empirical investigation.

Huh used ChatGPT to answer questions from a parasitology exam taken by first year medical students [24]. The answers were then compared to those from a corresponding cohort of students (n = 77). ChatGPT scored 60.8% compared to a mean of 90.8% for the students (minimum score 89.0%, maximum score 93.7%). The authors suggest that the high performance of the students may be due to the exam being sat shortly after the end of the corresponding module, and that long-term scores would likely be lower. They also point out that ChatGPT's inability to interpret figures, graphs, and tables made it reliant on the quality of the experimenters' textual descriptions. ChatGPT was also limited by the lack of Korean-language material and Korea-specific topics in its corpus. Finally, ChatGPT struggled with the formatting of certain multiple-choice questions.

Gilson used ChatGPT to answer questions on the United States Medical Licensing Examination Step 1 and Step 2 to qualitatively examine the integrity of its responses as a medical tutor [22]. The experimenters created two sets of data from well-established question banks (AMBOSS and NBME). Any questions reliant on understanding of an image were removed, as were those requiring the answer had to be formatted as a table. ChatGPT was prompted by providing the question text. Each question was asked twice: once with the simple question, the second time with a tip supplied from the original question bank.

ChatGPT performed better on the Step 1 questions for both AMBOSS and NBME than it did for Step 2. It also performed better on NMBE than AMBOSS. Performance decreased with increasing difficulty of questions. ChatGPT answered correctly more than 60% of questions at Step 1 and Step 2 (60% is considered a threshold pass for a 3rd year medical student). However, performance dropped steeply on AMBOSS without tips from 64% accuracy at Level 1 to 0.0% on Level 5. For AMBOSS Step 1, ChatGPT performed at the 30th percentile without tips, but increased performance to the 66th percentile with the tips. On Step 2 it went from 20th percentile to the 48th with tips.

In a rare example, Yeadon *et al.* outlined a study of ChatGPT's performance in essay-based assessment for an undergraduate physics curriculum at a prestigious UK university [51]. ChatGPT was used to generate 10 scripts for an exam comprised of 5 questions requiring short-form (300-word) answers. It was found that the AI responded in a more discursive manner when given richer prompts including limiting word counts or including a known historical figure or event in the question. Not only was ChatGPT capable of quickly generating synthetic solutions, but these solutions received an average mark of 71 ± 2%, with grades more tightly grouped around the average than in student populations. The researchers report that students performing in the lower-third of their cohort would improve their grade if they relied on ChatGPT for their solutions. Unsurprisingly, given that the software generates novel text rather than copying it from elsewhere, the scripts received extremely low scores from the anti-plagiarism software.

An experiment with 18 undergraduate forensic science students found no evidence that ChatGPT helped students outperform peers with essay writing tasks [1]. The cohort was split into two, with each group asked to answer the same essay question; one group had access to ChatGPT, and the other wrote essays using 'traditional' resources under supervision to ensure that they could not use the chatbot. The essays were double marked as well as being checked by an AI classifier to attempt to determine their origins. The classifier determined that 3 of the ChatGPT essays were of "possible" AI origin and 5 were "likely"; in contrast, 2 of the control group essays were "possible" and one was "likely". Not only did the control group outperform those students who used ChatGPT, but those students whose essays were flagged as of "likely" AI origin received lower marks. The experimenters suggest that lower grades for ChatGPT essays may be down to student unfamiliarity with the chatbot. This experiment aligns broadly with an earlier study by Fyfe in which students reported that incorporating ChatGPT outputs into essays was harder than writing the essay from scratch and did not represent a significant time saving [20].

2.5 Detecting LLM-enabled cheating

Significant effort has been invested into the detection of whether a given piece of text was written by a human, or by an LLM such as ChatGPT. Broadly speaking, these perform a statistical analysis of the text and provide a percentage likelihood of generation. A significant number of false-positives (that is: human-originated text being rated as synthetic) are reported across tools. Unlike plagiarism checkers, such as TurnItIn, which definitively link a student submission to a pre-existing piece of text, these tools cannot demonstrate "proof" that text has a synthetic origin – only the probability that it may be machine-generated, limiting their suitability in academic conduct cases.

Whilst there is little published with GPT-3/ChatGPT, there is a healthy body of literature of experiments using GPT-2 and other LLMs. Ippolito *et al.* noted that human discrimination of LLM outputs is generally lower than software detection, and that, to disguise their origins, LLMs tend to add statistical anomalies that potentially allow consistent detection of artificial solutions [26]. Rodriguez *et al.* note: *"it is significantly harder to detect small amounts of generated text intermingled amongst real text"* [37]. This is an obvious challenge, given that it is likely that most students using LLMs to cheat would use their outputs to supplement their own work rather than relying on an LLM to produce the entire submission. The synthetic origins of text will inevitably be further disguised by students rewriting and manipulating the material or employing tools such as grammar and style checkers built-in to most modern word processors, or external grammar tools. Sophisticated cheats could use the same online tools used by universities for the detection of synthetic texts to identify incriminating text in their submissions and make edits to change the suspicious content before submission.

Those ChatGPT detectors that have been released – including GPTZero, TurnItIn, and the Originality.AI detector – have yet to be independently assessed.

2.6 Research Gap

This literature review highlights how cheating has long been an issue. ChatGPT poses specific issues in terms of convenience and cost, as well as the generative nature making traditional plagiarism software inadequate. Further work is needed to assess empirically both the capabilities of ChatGPT and the risks it poses to academic quality assurance. This led to our two key research questions:

RQ1: How well does ChatGPT perform on diverse CS assessment material?
RQ2: Can experienced ALs distinguish between genuine student solutions and those generated by ChatGPT?

3 METHODOLOGY

To investigate these research questions, we designed a blind study protocol focused on providing synthetic and student scripts to volunteer ALs who were asked to mark the work as part of a "quality assurance" study. Full details on the methodology are outlined in this section.

Several modules from across the undergraduate and postgraduate Computing programmes were selected, to explore ChatGPT's ability to generate text at various levels of academic attainment and for different audiences. These modules use a range of assessment methods, requiring students to demonstrate a range of academic skills when constructing solutions. For the purposes of this study, we restricted our work to the final assessment, be that an exam or EMA. The selected modules are outlined in Table 1.

Table 1: Details on the modules used for the study.

Module	Level	Topics	Assessment	Format of assessment used in study
TM112	Introductory	Python programming, computer architectures, cloud	Coursework at three points in the module. The third, and final, assessment element was used for this study, and contributes 50% of the module grade.	Seven questions, including short definition-based questions, a programming exercise, several short essays (up to 500 words) and a reflective component based on the student's experience of the module.
TM129	Introductory	Samplers in computer networking, operating systems and robotics.	End-of-module assessment (EMA).	The EMA is designed to evaluate each of the samplers, with three short discursive questions and three longer essay questions requiring integration of knowledge and cited research. Essay templates are provided for the longer answers, of which the students answer any two of three. The final element is a reflective PDP (personal development planning) component.
TM254	Intermediate	Service management, project management, requirements and databases.	Exam.	Ten questions exploring aspects of database design and service management. The questions are tightly constrained and most contain multiple short sections. Several questions are based on case studies outlined in the assignment itself and require students to apply their existing knowledge. In one question, students are asked to write a series of SQL queries to perform specified tasks.
M269	Intermediate	Algorithm design and programming processes.	EMA	Seven questions examining the algorithm design process as well as program development. One question requires students to explain fundamental

Module	Level	Topics	Assessment	Format of assessment used in study
				concepts in computing whilst another examines proficiency in predicate logic. An essay question (maximum 800 words) requires students to discuss computability for a general non-specialist audience. The format of the essay is tightly constrained.
TM356	Advanced	Interaction design.	Open book exam.	15 short and one long question. Many of the short questions require students to briefly explain named concepts in interaction design in the context of specific scenarios. The long question concerns the development of a health tracking application for senior citizens. Students create suitable questions that could be used in semi-structured interviews in the elicitation of requirements and to prototype an interface for the application answering a defined specification
M811	Postgraduate	Information security through the lens of the ISO27000 family of standards.	EMA.	An extended report exploring a topic of the student's choosing related to a case-study developed during the module. As appropriate for a postgraduate module, the report is expected to be written to a high standard in 'academic language' and to employ extensive referencing of module materials as well as external resources.

3.1 Selecting and generating scripts

The study used two forms of solutions:

1. those previously submitted by students, and
2. 'synthetic solutions' generated using ChatGPT.

3.1.1 Student solutions

The research team was given time-limited access to the University's assessment archive. For each module, ten student solutions were randomly selected from a cohort that commenced study in late 2021 (coded **21J/K** below). This cohort was chosen not only because all grades had been finalised and any appeals processes completed; but because it was the last full cohort to finish studies before the public release of ChatGPT, meaning that ChatGPT could not have been used to create the solutions. Whilst student solutions from older presentations were available, changes to module content and assessment approaches potentially meant that those documents were not representative of current teaching practice. This set of assessments was the first to have been written since the introduction of at-home examinations as a response to the Covid-19 pandemic.

3.1.2 Synthetic solutions

ChatGPT was used to generate five synthetic solutions for each module. Five postgraduate student volunteers each produced all the synthetic solutions for a given undergraduate module, with the M811 scripts being generated by the research team. The instructions provided to the students noted that the chatbot should be prompted using the original assessment questions in a single ChatGPT thread. An extended dialogue between the user and chatbot was required for questions consisting of multiple sections. If further detail was required, the chatbot would be requested to 'continue' or repeat its explanation using alternative wording.

The chatbot's outputs were copied into pre-prepared Microsoft Word template documents which had random styling applied to replicate the diversity of student responses. The volunteers generating the synthetic solutions were not allowed to add their own text, edit responses, or remove irrelevant material other than to remove any evidence the text had been generated by ChatGPT.

The synthetic solutions were scrutinised by a second person to ensure the removal any evidence that they had been created using ChatGPT. We were aware that this is the most naïve way of generating material through ChatGPT, not accounting for students rephrasing material, adapting questions to account for image-based questions which couldn't be answered; providing additional prompts, or augmenting additional content. Our results therefore represent a baseline of minimal-effort cheating using ChatGPT.

The volunteers preparing synthetic solutions were asked to record how long it took to generate each solution, as well as any problems they encountered. During this study, ChatGPT reported significant capacity issues resulting in slow performance and occasional periods of complete unavailability, meaning the times taken are at the upper end of the technology's capability.

Table 2: Mean time taken to generate synthetic solutions using ChatGPT.

Module	Mean time (minutes)
TM112	23
TM129	15
TM254	32
M269	34
TM356	32
M811	25

Table 2 shows the mean time taken to generate synthetic solutions for the six modules. There was only one occasion on which script generation took more than an hour; but even this case would have allowed a student to complete the solution within the time allocated for an examination (usually two or three hours). For modules using EMAs, students are allowed several weeks to develop their solutions, meaning that even prolonged periods engaging with ChatGPT, or completing their submissions over several sessions, is feasible within the deadline.

Most of the volunteers reported minor problems whilst generating synthetic solutions; the majority were caused by capacity issues on the OpenAI site. The experiment coincided with intense public interest in ChatGPT, and it was not uncommon to be blocked from the service for prolonged periods of time. These constraints have continued, albeit less commonly; however, the paid-for ‘premium’ service promises much better availability.

3.2 Anonymising documents

Every effort was taken to anonymise student solution documents. Copies of the selected scripts were downloaded to a secure system as Microsoft Word documents. The scripts were anonymised and checked by two members of the research team to remove any identifying information (such as personal data or references to employers and sponsors) and then assigned new, unique identifiers.

Identifying metadata was removed from student and synthetic solutions alike. This not only removed author information, but reset the document creation and save dates, as well as erasing the document editing history. Had this metadata been retained, it would have been relatively simple for markers to distinguish genuine work (written by students over extended periods during 2022) from AI-generated synthetic solutions (created by the project team over a short period in early 2023 after the end of the module presentations).

3.3 Markers

Volunteer markers were recruited from the existing AL community. We were unable to recruit our target of two ALs per module; in some cases, we could not recruit any markers for a given module. TM112, TM129, TM356 and M811 were marked, with TM129 and TM356 scripts being independently double marked, (a routine process performed by the OU to ensure consistency between markers), identified below as markers TM129_1 and TM129_2, and TM356_1 and TM356_2 respectively. The markers were paid for two days’ work, using the OU day-rate equivalent for ALs.

3.4 Marking

For each module, the ten anonymised student solution scripts were mixed with the five synthetic solutions and given filenames from 1 to 15. Ordering was randomised, with the exception that, in any batch of solutions, one pair of synthetic solutions was always given consecutive numbers (*e.g.*, scripts 6 and 7). Solution documents were copied to markers’ personal confidential Microsoft Teams channels created for this study. Markers were then notified that the solution documents were available for marking.

Markers were asked to provide feedback through a marking table (simulating an existing feedback system), where ALs entered marks. They were also able to highlight suspected plagiarism or questionable content; recommend remarking; or identify problems with the script itself. ALs marked the scripts using pre-existing marking guides developed by the module teams. Markers were given up to three weeks to complete the task.

The markers were able to provide general feedback to the research team about the ‘quality’ of an individual script. Guidance to markers about this feedback was deliberately vague, to avoid alerting them to the research being performed; markers were simply asked to report *“anything unusual or unexpected found in the scripts”.* After completing marking, the ALs returned annotated scripts and feedback forms to their Microsoft Teams channels.

We concluded by inviting our markers to participate in a short semi-structured interview which was audio and video recorded. Five of our six participants took part. The interview script focussed on individuals’ marking processes, and how they typically detect and engage with academic conduct issues. After asking the markers what they thought the purpose of the marking exercise had been, further questions explored participants’ observations about student work, perception of ChatGPT, and their familiarity with the tool.

3.5 Analysis Techniques

A mix of descriptive statistics and graphing of the data was used to identify trends and patterns amongst the marking behaviour of the ALs (in terms of marks awarded). The interviews were audio recorded and transcribed. An inductive open coding approach was used to identify concepts and themes within the interview transcripts [6]. The transcripts were subjected to a line-by-line analysis by the first author, who had not interviewed any of the participants. Through this initial analysis, concepts were identified and labelled within the data. No codes existed prior to the analysis; they were created through constant comparison of the data and the application of labels to the text. This process was tempered by our interest in our research questions. These codes were subsequently categorised into unifying themes by the first author. These themes were then discussed with the interviewer, to ensure that the developed themes corresponded to her interpretation of the data.

3.6 Study perception

Table 3 summarises the marker’s perceived purpose of the study, and their experience with ChatGPT. It highlights a range of experience levels with ChatGPT, and that we were broadly successful in masking the purpose of the study.

Table 3: Participant breakdown for the interviews. TM356_2 declined the invitation to be interviewed.

Marker	Interview length (minutes)	Perceived study purpose	Experience with ChatGPT
TM112_1	27	Identification of “strange” scripts	Not used, aware from the news
TM129_1	41	Feedback on marking processes	Aware but not used
TM129_2	42	Consistency of marking	Has tried with assessment material
TM356_1	35	Consistency of marking	Aware but not used
M811_1	23	Chatbots	Aware but not used

3.7 Data protection and ethical considerations

The project team engaged fully with the OU's data protection and ethics policies, gaining appropriate permission from the institutional committees.

Since this study was solely for research purposes and was kept separate from formal module presentation and assessment procedures, the official grade for a student's work would not be altered if a re-marked assignment was awarded a lower grade than on first marking. Likewise, disciplinary procedures would not be invoked if re-marking produced evidence of plagiarism.

Before starting the marking exercise, our volunteer markers completed a consent form. They confirmed that they understood that the interview was optional, but by confirming participation in the exercise, they would complete the allocated marking duties. If they opted to be interviewed, markers consented for it to be recorded and anonymous quotes to be used in publications. We allowed participants the right to withdraw their interview data within two weeks of the interview; none of the ALs did so. The consent form also confirmed that the marking exercise data could be used for academic publications in an anonymised form, and that they would keep the content and marking of scripts confidential.

We committed to informing the ALs of the true purpose of the study, either at interview or, for those ALs who did not want to be interviewed, by email.

4 MARKING EXERCISE RESULTS

Driven by the two key research questions (RQ1: How well does ChatGPT perform on diverse CS assessment material? and RQ2: Can experienced ALs distinguish between genuine student solutions and those generated by ChatGPT?), our results are presented in four sub-sections. The first subsection covers the performance of the synthetic solutions in the blind marking by ALs, whilst the second subsection examines the identification of synthetic solutions by ALs. Subsequent subsections consider the key plagiarism practices identified by the ALs in interviews, before examining practices regarding referencing in detail.

4.1 Marking Performance

The student and synthetic solutions for four modules were marked by ALs, with TM129 and TM356 being independently double marked. The final marks after any deductions are shown below in Table 4 and Figure 1.

Table 4: Final percentage scores for synthetic and student solutions. Synthetic solutions are grouped at the top of the table, student solutions are at the bottom.

Script type	TM112_1	TM129_1	TM129_2	TM356_1	TM356_2	M811_1
Synthetic	80	67	54	63	51	13
Synthetic	83	69	65	66	50	21
Synthetic	84	60	66	66	44	36
Synthetic	86	67	71	67	45	41

Script type	TM112_1	TM129_1	TM129_2	TM356_1	TM356_2	M811_1
Synthetic	86	52	55	66	46	16
Student	52	97	85	71	54	88
Student	91	84	70.5	86	82	40
Student	49	78	48	46	29	53
Student	95	88	78	54	39	20
Student	72	80	74	67	45	0
Student	70	94	75	66	63	0
Student	48	86	75.5	79	80	36
Student	63	82	76	79	68	62
Student	95	86	79	53	42	46
Student	74	60	49	75	64	47

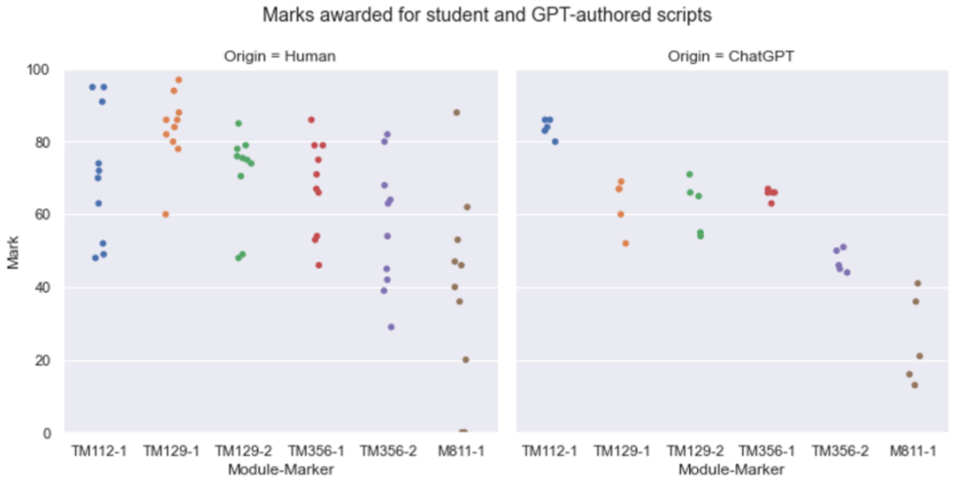

Figure 1: Comparative marks scored by students (left) and synthetic solutions (right).

In the case of TM112 and TM129, every single synthetic solution achieved *at least* a passing grade. For TM356, one marker awarded a failing grade (<40%) to two synthetic solutions; the same solutions were awarded a passing grade by the second marker. Every synthetic solution for M811 was marked as a *'fail'* (postgraduate modules at the OU have a higher pass threshold of 50%).

Within the passing grade, the OU can also award a 'distinction' grade for especially high-scoring student solutions. This is typically awarded for scripts scoring greater than 85%. Distinction grades would have been awarded to three of the TM112 synthetic solutions and one synthetic solution for TM129. These results indicate that a student wishing to cheat by using ChatGPT to generate the entirety of their assignment solutions could expect to pass these end assessments for both TM112 and TM129 and would have a high probability of passing TM356.

A key part of module assessment is the determination of final grades. Realising that any form of marking involves subjectivity on the part of a marker, module teams may choose to re-mark solutions just below grade boundaries to decide whether a higher grade is deserved. Therefore, solution documents lying close to these boundaries receive greater scrutiny. Since most synthetic solutions received scores away from boundary grades, had they been submitted as genuine assignment documents they would not be further scrutinised before being awarded a passing grade. Perhaps the most significant finding is not that ChatGPT behaves as an outstanding student across some of the undergraduate modules, but that it performs consistently as an ‘adequate student’ – able to pass assessment without drawing undue attention to itself.

4.1.1 Comparison of the script sample against cohort norms

Having established that most synthetic solutions received pass marks, it is worth examining how the sample scripts – both synthetic and student – compared to the wider cohort. Table 5 shows that the mean and standard deviations for marks align broadly with mean marks for the wider cohort. Except for M811, mean synthetic and student solution scores lie within one standard deviation of the cohort mean. For TM112, the synthetic solutions outperformed both the cohort and student sample; whilst ChatGPT underperformed relative to both for TM129 and TM356.

M811 represents an outlier; not only did ChatGPT perform poorly in the assessment, but the student sample is unrepresentative of the wider M811 cohort. This does not affect the validity of the results for the synthetic M811 solution documents, but it does prevent broader comparisons against the performance of the scripts against the cohort as a whole.

Table 5: Performance of the synthetic and student solutions in the experiment against that of the entire 21J/K cohort for the selected modules.

	21J/K cohort			Synthetic		Student	
Module	Cohort size*	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
TM112	1,317	79.03	16.85	83.8	2.49	70.9	18.26
TM129	925	73.7	20.13	62.6	6.82	77.25	12.77
TM356	195	64.91	20.5	56.4	9.97	62.1	16.19
M811	73	62.62	11.05	25.4	12.42	39.2	27.19

* Cohort size represents the number of students submitting their final assessment.

Table 6 shows the spread of marks awarded for synthetic and student solutions.

Table 6: Spread of total marks awarded to synthetic solutions and students for each module.

Marker	Synthetic solutions	Student solutions
TM112_1	6.00	47.00
TM129_1	13.08	28.46
TM129_2	13.08	28.46
TM356_1	4.00	40.00
TM356_2	7.00	53.00

Marker	Synthetic solutions	Student solutions
M811_1	28.00	88.00

In each case, the synthetic solutions demonstrate a lower spread of marks than the genuine student solutions. Since each of a module's synthetic solutions were generated from identical prompts, there was limited scope for ChatGPT to generate radically different outputs. As mentioned earlier, we did not edit the output of ChatGPT when creating the synthetic solutions, so this low level of mark spread may not be representative of actual behaviour from cheating students.

The low level of mark spread is interesting as it demonstrates that the generation of answers through ChatGPT produces extremely similar content. This is unsurprising to some extent given that LLMs lack conceptual knowledge and are "stochastic parrots" [2].

Again, M811 represents an outlier. The extremely large spread in student solutions on the module is, in good part, due to two student scripts being awarded zero by the marker. In one of these cases, the student had erroneously submitted a solution to a different piece of assessment; in the second, the researchers believe the solution was erroneously marked as zero and should have been awarded approximately 15%.

4.1.2 ChatGPT performance on individual questions

We wanted to examine the marks awarded to individual questions to identify strengths and weaknesses in terms of the nature of topics being assessed. To do this, we categorised the questions by topic. We then examined the marks and colour-coded the question into six bands:

1. All scripts received a distinction mark for this question.
2. A majority of scripts received a distinction mark for this question.
3. Neutral performance.
4. A majority of scripts received a fail mark for this question.
5. All scripts received a fail mark for this question.
6. Not attempted.

Please see Appendix A.1 for a complete breakdown for all questions in TM112, TM129 and TM356.

Based on this analysis, we loosely identified some trends. For TM112, questions requiring a discussion of program development improvements and definition questions regarding the usage/vulnerabilities of SQL all received a distinction mark. For the questions focussed on security and hashing, which involved stating definitions and applying techniques to simple scenarios, a majority of scripts received a distinction mark. The lowest scoring questions – receiving neutral marks – assessed programming and program development; an essay on social, legal, and ethical issues around digital literacy; and reflecting on module performance.

For TM129, there is mixed performance across the questions. The essay question on the use of robotics in space exploration, including sourcing examples with references, was answered well, with a majority of scripts receiving a distinction mark. The essay question on operating systems, building on a video on the history of Unix, received neutral marks. Broadly, the three short-answer opinion questions on O/S, robotics and networking also received neutral marks.

The PDP questions were mostly handled poorly, with all the ePortoflio question responses failing; the majority of the self-reflection questions failing; and the future planning tending to receive neutral marks. We did not use one of the optional long-form questions, as it was deemed unlikely that ChatGPT would be able to handle the practical networking activity using a network simulator without a huge amount of editorial work around the questions.

For TM356, there is a significant consistent discrepancy between the two markers. That said, all 15 of the short-form questions requiring students to apply their understanding of a key HCI concept – from requirement gathering to design techniques and evaluation – broadly received neutral or higher marks from one marker, and majority distinction or higher from the other maker. For the long-form scenario question, the first part, which covers the development of interview questions, received neutral or higher marks. The second part, requiring a heuristic evaluation of several interface screenshots, could not be passed through ChatGPT. The third and final part, requiring a redesign of an interface building on the heuristic evaluation, received neutral or lower marks.

While it is obviously challenging for any given question to identify whether the question format or the topic covered was responsible for the ability – or not - of ChatGPT to generate compelling answers, this analysis demonstrates that, in most cases, across a range of question formats, topics, and study levels, ChatGPT is *at least* capable of producing adequate solutions.

4.2 Synthetic script identification

One of the most significant observations from the study was the distinction between ALs' ability to recognise suspect scripts, and their decisions about which scripts to flag to the university. Although the number of scripts flagged as 'suspected of plagiarism' by the ALs was small, at interview it was clear that their identification of synthetic solutions was more accurate than the flags alone might suggest.

Table 7 shows the plagiarism flags that the markers entered in the formal marking table during the marking exercise. An additional 9 scripts were identified during the interview as appearing "suspicious" without meeting the bar for plagiarism flagging. This is discussed in depth in Section 4.3.

Table 7: The number of scripts flagged by ALs as being suspected of plagiarism. *

Flag type	Synthetic solutions	Student solutions
Plagiarism flag during marking	7 [9 additional scripts flagged at interview]	3
Not flagged	23	57

Of the three flags on student solutions, one refers to a script which is currently under active investigation by the Academic Conduct Officers and the OU; this is one of two such cases among the scripts sent to markers. Both of the other scripts flagged by markers referred to scripts where two of the three short-form questions (each worth 5/100 marks) either rely on re-writing an external source or make too much use of quoted material. Neither of these scripts would likely be referred for an ACO investigation, given that the two long questions (30 marks each) contained no academic conduct issues, and the scripts scored 18% and 16% respectively on TurnItIn. Overall, the ALs were effective in detecting plagiarism in student scripts.

The ChatGPT scripts flagged on the marking table were identified by the markers for the following reasons:

- Marker TM112_1 commented on one synthetic script that *“Parts of the Questions 6 and 7 appear to be missing.”* and on a different script that *“[this script] and [the other flagged script] are similar for Qu. 2(ii) for example.”*
- Marker TM129_1 noted on one synthetic solution: *“A mixture of some very good elements, and some less good. Inconsistent across questions. In a TMA, this would raise concerns for me.”*
- Marker TM129_2 identified four distinct scripts from TM129_1, which shared a common, unusual approach to answering the same question, commenting *“Strange angle to answer - would swear it came from somewhere”* particularly in reference to the personal reflection element of the exam. These four scripts were identified as being unusual by TM129 AL1, but the marker did not feel they met the threshold to raise a plagiarism flag.

From the in-script comments left by these markers, it is clear in each case that the questions raising concerns require students to provide either a personal viewpoint or some self-reflection. The answers provided are either missing that sense of personal reflection (as with TM112), or a synthetic solution provides an answer that is written in the third person, talks about reflection in broad terms, and doesn’t relate back to an individual experience. The next section builds on this analysis by considering the data from the interviews conducted with ALs, examining their practices regarding academic misconduct more broadly.

At the end of the interview, we informed the ALs that five of the marked scripts had been authored by ChatGPT. When participants were asked if they were surprised that some of the scripts had been generated by ChatGPT, none of the markers were surprised. The ALs were asked to revisit their identification of synthetic solutions. All bar one of the markers re-identified scripts they had previously flagged or identified as suspicious, with the same justification. The two TM129 markers had, between them, flagged all 5 synthetic scripts; at interview, TM129_1 increased identification from 1 that had been flagged to all 5. M811_1 hadn’t flagged any of the scripts for plagiarism in the marking feedback, but immediately identified three of the synthetic solutions (1, 10, and 11) and on reflection identified the two others (5 and 15). M811_1 also raised flags on some of the student scripts, likely due in part to the selection of student scripts not being reflective of the typical range (see Section 4.1.1). The only TM356 marker we interviewed – TM356_1 - responded by discussing suspicious symptoms, rather than identifying specific scripts.

Overall, ALs’ awareness of suspicious scripts, and their articulation of the characteristics that prompted suspicion, were impressive, although their formal flagging of scripts was extremely selective.

4.3 Interview analysis

We asked the interviewees about their practices regarding academic conduct: how they identified scripts of concern, and how they decided whether to flag them. Four of the markers (TM112_1; TM129_1; TM129_2; TM356_1) highlighted a general expectation that plagiarism would be detected either by central members of the module team or the University’s anti-plagiarism software. TM129_2 wrote: "*you always say to yourself why should I flag it up? Because it goes through three or four pieces of software anyway, and they'll flag it up*". This is a perfectly justifiable position, given a potential lack of understanding of central processes; limited time available for marking, and the range of different academic conduct issues that could be investigated: "*there's so much out there I could spend a lot of time looking*" (TM129_2).

Markers TM129_2 and TM356_1 both highlighted that they tended to have an extremely high threshold for academic conduct, seeing such situations as an opportunity for teaching and developing students. TM356_1 commented: "*I haven't flagged it with anybody else unless it's blatant. I tend to actually put it to the students, you need to… put this in as a reference. And this means not just in the references at the bottom, but in-text citations to say that this is where I got this information from, and I tend to flag it to the student in that way*". One of the reasons for doing so was provided by TM129_2: a wariness of mislabelling student work, both due to the impact on the student, and *potential* consequences for the marker themselves.

Given the close relationship between students and markers, it is perhaps unsurprising – particularly in circumstances where marking time is limited, and robust procedures are in place for other module team members to assess misconduct – that markers focus principally on *teaching*, i.e., on helping students improve their practices, rather than defaulting to disciplinary procedures, unless the misconduct is both blatant and severe. While this is an exemplary practice, it is unclear how well it will serve in the new era that ChatGPT has created.

The identification of potential cases of academic misconduct is especially difficult in a distance setting, where "*there's a fine line between collaboration, peer learning, and collusion. And that's an interesting challenge*" [TM129_1]. Given that much of the correspondence and interaction between students, and between students and staff, occurs asynchronously and typically online, it can be challenging to work out where to draw a boundary between acceptably collaborative learning and outright plagiarism. Current practice – as noted previously – tends toward providing study skills support, rather than activating disciplinary procedures, although this did depend partially on the module level.

The ALs were very clear in distinguishing between aspects of student submissions that served as flags of misconduct and aspects that were the result of typical student behaviour.

The most-mentioned flag was a change in style in the answer. This could be differences in the layout of the document itself or changes in the use of language, such as: changes in tone or voice; the use of technical vocabulary; the proficiency of writing; or the specificity of the answers given: "*obviously if you're reading through something and there are significant language differences between answers to different questions*" [TM356_1]. This was particularly the case with changes in the use of technical language: "[where previous answers were] *very general and didn't know any technical detail, and suddenly you get an answer that's full of technical detail, and something like that makes you very suspicious*" [TM129_2].

This sense of consistency as a key indicator to authenticity led some tutors to choose to mark complete scripts in order to perceive the student voice – rather than marking each question in turn across all scripts. Similarly, in their usual marking practice, at least four of the markers looked back at previously submitted work, as "*you can spot that there's something really going wrong, this student is totally, totally different*" [TM129_2]. However, as TM129_1 noted, while consistency matters, sometimes it's hard to identify issues as "*they've just answered these questions on different days and didn't proofread*".

TM112_1 noted that many of the stylistic flags were more pronounced in material requiring students to integrate external material, noting that – particularly at Level 1: "*if they're asked to read a paper and they don't actually put things sufficiently into their own words…*".

As the final flag, M811_1 noted that repetition acts as a significant flag: “*it was the repetition that set me off*". Given the long-form essay style of this question in which a strong narrative element is expected, it is perhaps unsurprising that such a behavioural pattern, typified by ChatGPT’s responses, acts as a clear flag requiring further investigation.

The ALs also identified student behaviours which were not flags for misconduct but were accepted as baseline behaviours requiring study support. In introductory modules, tutors are more forgiving as “*I think their quality of academic writing tends to be flakier because they just don't have the experience in it*” [TM129_1].

TM356_1 identified the key points of assessment, namely: “*One is, do you know what you're talking about? Two is can you apply it? Three is can you communicate that information in an effective way?*”. This was deemed challenging to use as a flag, as students display various sub-optimal behaviours. These include:

- Not answering the question: *`ohh, it's about this. I'll just tell you everything I know about this’. Especially in an exam situation; you're more likely to do that rather than stop and think*” [TM356_1].
- Not showing working: “*if they’ve done a calculation and they have bothered to put any working in, and then suddenly there’s one with lot of working”* [TM129_1]
- Not applying the scenario context to the answer given: “*I did find a lot of them weren't really utilising the scenario, the context they were given*” [TM356_1]
- Student performance being extremely variable: "*the quality of students’ work is often quite variable*" [TM129_1]

Recognising that these are familiar student behaviours is limiting, as script feedback for the synthetic solutions shows similar patterns of behaviour. Hence, these behavioural cues cannot be used to distinguish between student scripts and synthetic solutions without additional evidence.

4.4 Referencing

ChatGPT is known to it generate false references [50]. This is unsurprising as – in its current form – ChatGPT is probabilistically generating text rather than performing searches. We leave to one-side whether future versions, with live links to services such as Google Scholar or PubMed, could replicate real references, without any understanding of the content of the underlying paper.

Recognising this aspect of ChatGPT, we examined every reference provided in the synthetic solutions. All of the references that were included were either artificial or drew on material in the question used in the original prompt.

This behaviour is best illustrated by the M811 EMA, which consists of long-form essay based on a critical review of three recent security-related research articles. The 5 synthetic solutions contained 26 references, all of which were incorrect. 25 referred to genuine journals and one to a conference, all of which were relevant and with correct volume numbers and dates. 12 of the page numbers were found inside the named issue but didn’t refer to a paper, while 14 lay outside of the issue named in the reference. All but one of these references referred to a non-existent paper title. 7 of the references named actual authors who have published in the area of security. 18 included no URL or DOI, whilst 5 included an invalid URL/DOI, and 3 included a URL/DOI to a different paper.

For example, the reference below is formatted correctly according to OU practice. However, whilst the authors have previous collaborated on security papers, this paper title doesn't exist. The name of the journal, dates and volume and issue information are all plausible, but the page numbers lie beyond the last page of the named issue.

> Title: "Adversarial Examples in the Physical World: Lessons Learned and Challenges Ahead" Authors: Battista Biggio, Fabio Roli, Blaine Nelson Publication: IEEE Security & Privacy, vol. 17, no. 3, pp. 84-92, May/June 2019. Link: https://ieeexplore.ieee.org/abstract/document/8706763

The plausibility of the references is remarkable. At first sight, the journals are relevant, they broadly have the correct form, and the authors are sometimes sensible – there are few flags indicating that these are nonsense references when reading at a surface level under time-pressure.

Given the known ability of ChatGPT to generate artificial references, we specifically asked participants about their behaviour around checking references,. Referencing has greater prominence in some modules (e.g., M811 requires in-depth discussion of student-selected references; TM356 requires no references). Unsurprisingly, ALs pay greater or less attention to referencing based on the module they are marking. Both TM356_1 and TM129_1 noted that they typically scanned the formatting and venue of the reference for correctness but didn't check them for validity as the time for marking is so tight. TM129_2 was similar but noted that they would follow the references if they suspected plagiarism. Both TM112_1 and M811_1 said that they would follow references, with the M811 tutor noting that based on the module material "*if you can't find a link to them, you can't identify them, then it's an automatic fail*". This helps account for the low scores received by many of the marked scripts, and why it didn't result in the scripts being flagged for plagiarism.

5 ACADEMIC MISCONDUCT SOFTWARE RESULTS

In addition to the marking exercise, we also examined how two key pieces of software coped with our script sample (including TM254 and M269).

5.1 TurnItIn

The TurnItIn service is applied to all assignments submitted to the Open University. TurnItIn identifies similarities with known sources of content including academic databases, websites and existing student submissions. The results can be used to identify cases of suspected plagiarism – with the proviso that submissions containing large volumes of quoted text may be flagged as plagiarism even when correctly referenced. Scores can range from 0.0 (no evidence of plagiarism) to 1.0 (material is entirely plagiarised). Module teams at the OU often consider a TurnItIn score of 0.25 or above worthy of investigation for further evidence of plagiarism. The TurnItIn results for our script sample are shown in Table 8.

Table 8: TurnItIn statistics for all solution documents used in this experiment.

	Synthetic solutions			Student solutions		
Module	Minimum	Maximum	Mean	Minimum	Maximum	Mean
TM112	0.12	0.21	0.16	0.06	0.62	0.18
TM129	0.05	0.08	0.06	0.03	0.23	0.13
TM254	0.02	0.10	0.04	0.00	0.03	0.01
M269	0.06	0.14	0.11	0.00	0.13	0.04
TM356	0.00	0.02	0.01	0.00	0.01	0.00
M811	0.01	0.18	0.09	0.09	0.74	0.26

Figure 2 shows TurnItIn scores for all 90 solution documents in the experiment. Synthetic solutions are grouped on the left of the figure with genuine student solutions on the right.

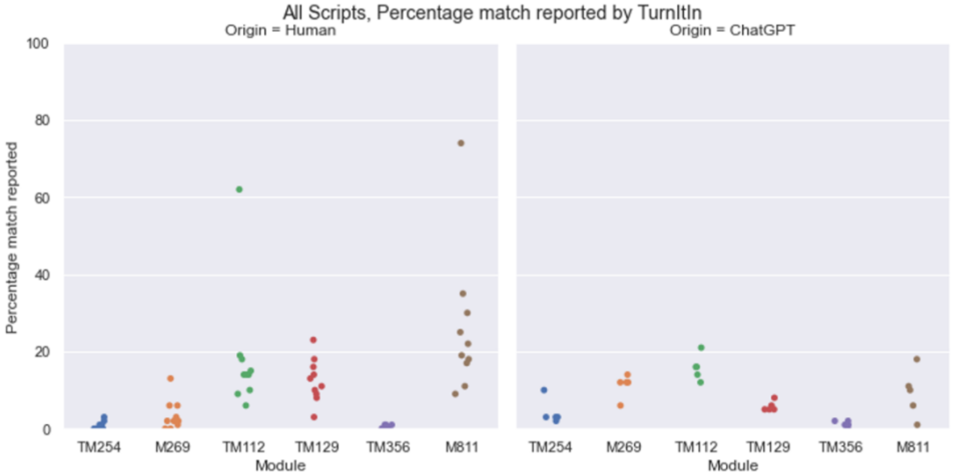

Figure 2 TurnItIn scores for all solution documents used in this experiment.

All synthetic solutions produce very low TurnItIn scores (mean = 0.079), indicating they are comprised of novel text rather than material sourced from elsewhere. The synthetic solutions have slightly lower mean scores than the selection of student solutions (mean = 0.102) as well as a smaller range of TurnItIn scores (0.21 compared to 0.74). Even when two anomalously high-scoring student scripts (TurnItIn scores of 0.62 and 0.74) are eliminated from consideration, ChatGPT consistently produced lower TurnItIn scores than genuine student scripts. This is consistent with the expectation that LLMs, such as ChatGPT, generate novel text rather than repeating existing material.

5.2 GPT-2 detection

Alongside the development of LLMs, software has been developed to identify whether a piece of text is human or LLM generated. This is typically performed by generating a statistical profile of the text, resulting in a numeric predictor of the likelihood of the text being synthetic. Unlike conventional plagiarism detection software, there is no demonstrable link to any evidence base indicating that the script has not been written by a student.

We wanted to explore how accurately the 90 scripts in our study (60 student scripts, 30 ChatGPT-generated) could be identified by detection software. We recognise that we have no absolute guarantee that *any* of the student scripts were *actually* written by a person, however, these scripts were received by the University prior to the public release of ChatGPT. On the balance of probability – given the relative complexities of using previous-generation LLMs – we believe that all 60 student scripts were written by people but acknowledge that this may not be the case.

Terms and conditions imposed by many ChatGPT detection systems pose significant ethical and intellectual property rights issues since these systems claim rights over all of the submitted data. The respective privacy policies give little detail how submitted data can be stored, used, shared or monetised; nor whether requests can be made for data deletion. Given the understandable data and ethical protections placed around student data by the Open University and UK legislation, the team chose not to submit the student scripts to any online ChatGPT detector.

Instead, the project team used a previous generation of detection software, (GPT-2: 1.5B (https://openai.com/research/gpt-2-1-5b-release), running locally. However, we concede that this detection software is likely to be less successful at identifying ChatGPT outputs than a dedicated GPT-3 detector.

The software allowed completed solution documents to be tagged by question number, providing a more detailed breakdown of potential areas of concern.

We ran all 90 solution documents through the detection software twice, each time in batches of five; the first pass with the solutions tagged question-by question, and the second pass running over the entire script. The results are mixed. Table 9 shows the percentage figures for the detection rates on scripts tagged with question labels whilst Table 10 shows the detection rate for entire scripts. We suspect a script of being of AI origin if it scored more than 0.2.

Table 9: Detection rates (%) for scripts tagged with question labels.

	Synthetic solutions		Student solutions	
Module	Flagged	Not flagged	Flagged	Not flagged
TM112	80	20	50	50
TM129	100	0	0	100
TM254	40	60	30	70
M269	100	0	30	70
TM356	100	0	40	60
M811	100	0	20	80

Table 10: Detection rates (%) for entire scripts.

	Synthetic solutions		Student solutions	
Module	Flagged	Not flagged	Flagged	Not flagged
TM112	0	100	10	90

	Synthetic solutions		Student solutions	
Module	Flagged	Not flagged	Flagged	Not flagged
TM129	20	80	0	100
TM254	0	100	20	80
M269	20	80	60	40
TM356	20	80	10	90
M811	100	0	20	80

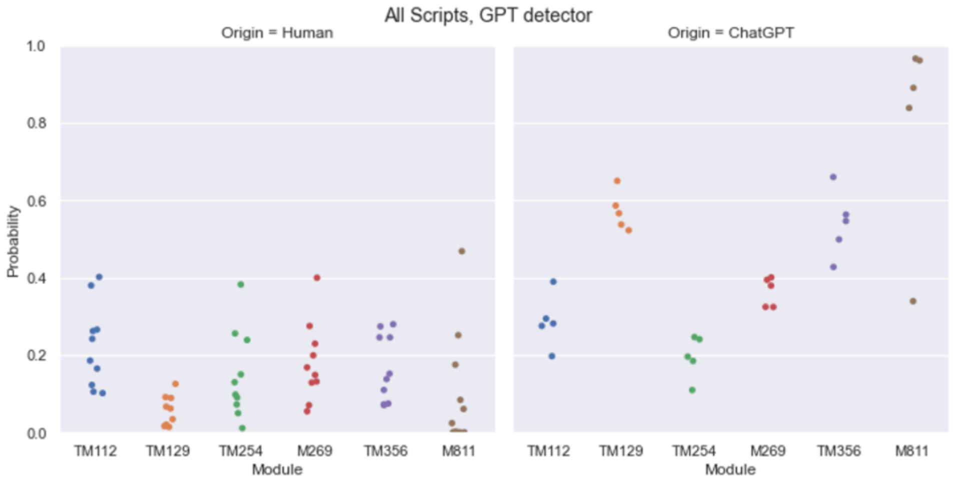

Figure 3 GPT-2 detector score for all solution documents used in this experiment.

Figure 3 shows the results of analysing all of the solution documents used in the experiment with the GPT-2 detector. The synthetic solutions have been grouped on the right of the chart with the genuine student solutions on the left.

40% of synthetic solutions score greater than 0.5 indicating they are more likely to have been generated by a LLM whilst no student solution scored above 0.5. The remaining 60% of the synthetic solutions score between 0.1 and 0.5 – very to 57% of student scripts. The remaining 43% of the student solutions lie in the band between 0.0 and 0.1 with no synthetic solutions scoring this low.

5.3 **TurnItIn AI detection**

In April 2023, the Open University chose not to opt-out of the AI writing detection system offered as part of its existing TurnItIn contract, instead choosing not to use any comparison data from the AI tool. Given the OU’s existing relationship with TurnItIn, it was possible to submit our sample set of 90 solutions for analysis.

Figure 4 shows that more recent detection tools trained on the GPT-3 and GPT-3.5 language models underpinning ChatGPT are far more successful at identifying ChatGPT generated material. The mean percentage of text in student solutions considered of AI-origin was 0.25% (SD of 1.17%, maximum of 8%). Comparatively, the mean for synthetic solutions was 74.43% (SD of 25.8%, minimum 28%).

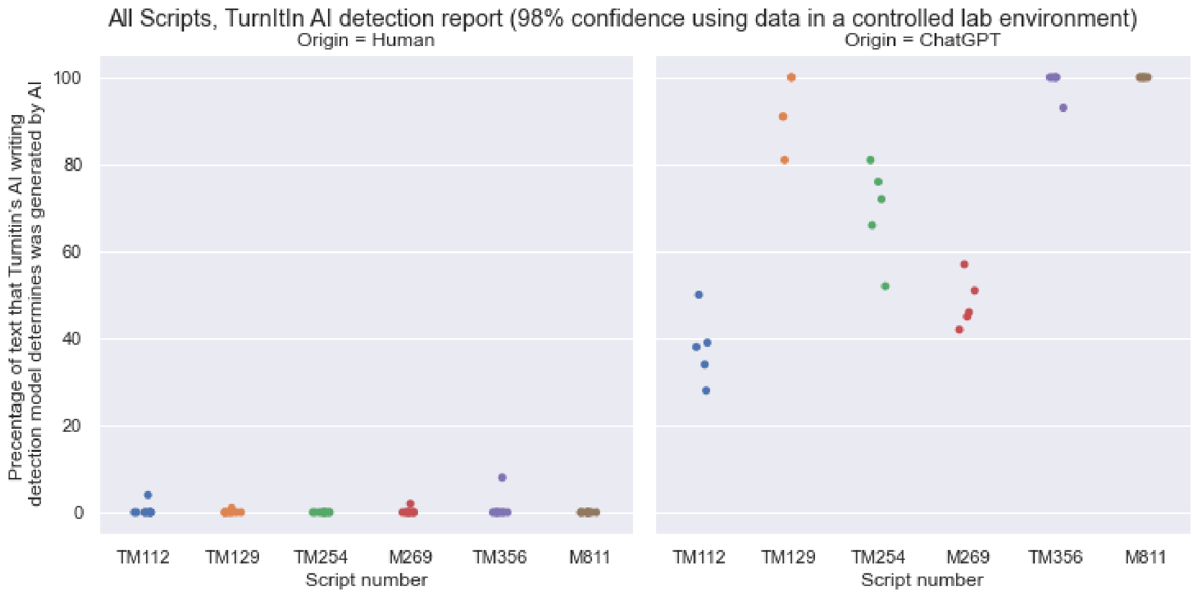

Figure 4 TurnItIn AI detection reports for student solutions (left) and synthetic solutions (right).

5.4 Analysis and Implications

Across the different tools, there are differences in their efficacy, with some producing a consistent 20-30% false-positive rate amongst student scripts, with other tools proving more robust. If a suitable detection tool was intended to select students for an oral examination, (the most likely use-case at the OU), these tools could prove helpful as a form of triaging candidates. However, given potential use of the tool involves legitimately use of generative AI tools to support disabilities and/or 2nd language development, institutions would need to carefully balance the biases that could be generated by using the tools for detection. Whilst the results from the TurnItIn AI detection software clearly distinguish student solutions from synthetic solutions, its creators warn that its results are purely interpretive and should not be used as the basis for bringing disciplinary cases against students.

Naturally, detection software cannot take account of other concerns around the use of ChatGPT including a student re-writing the material or a student passing the generated material through other software (such as Grammarly, or other text-adaption software) to rephrase the artificial text. As [37] notes, the integration of synthetic texts into a document containing some student-originated material is a common form of cheating that is much harder to detect through software.

In short, our findings can be summarised as "beware of snake oil" - technology alone cannot solve this problem. As the UK Quality Assurance Agency for Higher Education states:

"*Detection tools - be cautious in your use of tools that claim to detect text generated by AI and advise staff of the institutional position. The output from these tools is unverified and there is evidence that some text generated by AI evades detection. In addition, students may not have given permission to upload their work to these tools or agreed how their data will be stored.*" [35].

Whatever improvements are made to detection systems, some of the concerns raised – particularly the lack of concrete proof – indicates that detection software is not a panacea for quality assurance issues raised by ChatGPT.

6 DISCUSSION

There is a long history of resistance to new technologies in learning going back to Plato, through to the incorporation of calculators in the curriculum and the use of spell-checkers and word processors. The debate on digital technology in education has a range from those who think universal access to such technologies solves many problems through to those who think misinformed use of digital technology robs education of essential human values [47].

Our work clearly demonstrates that similar resistance is going to occur given the improvement in performance and ease of access of generative AI tools. Assessment material for varied standard topics from across the computing curriculum all appear to be answerable by ChatGPT.

This is perhaps unsurprising, given the early results we discussed in our literature review; there is nothing inherently different about computing topics that would make it harder to generate compelling answers from an LLM.

While investigating the performance of ChatGPT against a set of AP exams, Warner finds that the software is able to produce fluent, syntactically correct, well-structured answers that would probably be awarded a passing grade [48]. However, rather than be overly concerned by the AI's capabilities, they ask what such a result meant about existing assessment practices and whether academia can reasonably claim to be examining student proficiency when the necessary assessment can be passed by an AI that has no understanding of the subject domain.

Warner argues that universities have unwittingly developed assessment strategies that align with ChatGPT's strengths. Written academic assessment is often highly templated to support students who may be unfamiliar with writing large volumes of highly structured text. Guidelines often indicate the number and order of sections in a solution as well as their expected lengths; if referencing is expected, and if so – how many sources should be used, as well as the specific formatting of a bibliography; request users to define and use specific terms in their solution; or to address their answer to a specific audience. These well-meaning guidelines can help users create highly specific ChatGPT queries that produce apparently relevant results. Whilst removing such guidelines would make it more difficult for cheats to produce entire answers through ChatGPT, the downsides of doing so – disadvantaging weaker or novice learners – greatly outweigh the benefits.

Similarly, Warner pointed out that once calculators were introduced to the classroom they made many long-winded tasks simple and quick [49]. Whilst it is possible to forbid students from using calculators, the result is that they spend time performing identical mechanical tasks to the calculator - but take longer and are more error-prone. This time could be better used to practice the intellectual skills of mathematics. Akin to the role of the calculator, it is reasonable to ask if educators should continue to ask students to perform many tasks better suited to AIs.

6.1 Assessment design

Banning the use of generative AI is neither pedagogically sound, nor practical. Similarly, given that ChatGPT is accessible through any modern web browser; online proctoring is unlikely to prevent use of tools such as ChatGPT in an open book situation. Even where devices or browser software is locked down to prevent access, students could access ChatGPT on other devices such as smartphones or tablets. Proctoring is entirely unfeasible over extended periods of assessment such as project work, the development of a dissertation or the assembly of a portfolio that might extend over several months.

In terms of insights into how to structure assessment, ChatGPT performed relatively-poorly with questions requiring a high level of reflective content, or the assessment of group working. Similarly, our experiment showed that ChatGPT struggled to answer questions based around specific content not available outside of OU materials. This latter point may not remain relevant with the release of portable 'small LLMs' [43] that can be installed on personal computers and trained with specific data sets that could easily include OU materials.

Synthetic solutions answering questions that required finding, discussing and referencing academic publications were answered especially poorly. At the time of the experiment, ChatGPT was contained in a sandbox and did not have access to academic libraries. Consequently, it produced and confidently discussed fictitious publications in a superficial and repetitive manner as well as producing apparently convincing, yet entirely fabricated, references and bibliographies. Again, this may only be a temporary shortcoming as 'plug-ins' to ChatGPT will expand its abilities by drawing on external sources including academic journals; although it is unlikely that will improve its ability to discuss – rather than reference – those publications.

These recommendations echo those of Cotton et al., how argue future assessment should be made more resistant to LLMs [10], suggesting educators develop models featuring:

- Assignments demonstrating critical thinking, problem solving and communications skills;
- An increased requirement to supply credible citations and references;
- Open-ended assessment where students develop their own research questions as well as developing and defending arguments in an academic context;

Tate et al. propose a generalised pedagogy that could underpin subject-specific teaching to student raise awareness of and capability in using technologies such as ChatGPT [47]. It contains five key elements:

1. Understand. Providing students with a basic understanding of large language models as well as their strengths and weaknesses;
2. Access. The institution must provide students with access to the tools;
3. Prompt. Students need sufficient knowledge of the tools and critical thinking skills that they are capable of independently creating prompts that generate useful responses;
4. Corroborate. Students need to learn how to understand responses and then how to verify the quality and accuracy of the outputs, and;
5. Incorporate. Students must be able to include the outputs from these models in their own work, demonstrably adding their own value and correctly citing involvement of the model.

There is a desperate need to assess how computing curricula should design new assessment strategies that encourage appropriate use of generative AI (to prepare students for the future) while ensuring academic standards are maintained and is essential the best practice is rapidly disseminated.

6.2 Plagiarism detection

In addition to rethinking assessment, we also have to consider how best to prevent cheating. The results from detection software suggests that conventional automated screening processes cannot adequately distinguish synthetic solutions from genuine student solutions. The study shows that synthetic solutions show comparable – or even lower - plagiarism scores than student solutions, rendering anti-plagiarism software essentially useless. Dedicated GPT-2 detection software not only passes a significant number of synthetic solutions as student solutions (false negatives) but considers a substantial number of genuine student solutions to have been generated by AI – potentially leading to the triggering of disciplinary procedures against innocent students. Turnitin's AI software, trained on more recent GPT-3 and GPT-3.5 models was capable of distinguishing student solutions from those generated by ChatGPT; although, as its creators stress, the inability of GPT-detection to provide irrefutable evidence for the synthetic origin of text leaves any such determination open to appeal by students – whether or not they have actually cheated. Whilst it is reasonable to predict that detection software will become more effective; it is just as reasonable to assume LLMs will become increasingly proficient at generating outputs indistinguishable from naturally generated text. There is no panacea here; detection software is not going to be the best solution.

Honour codes can deter cheating and hold students accountable, although evidence of efficacy is limited. Punishment for misbehaviour – including cheating – demonstrates that there are consequences for dishonesty and serves as a disincentive to those who may be tempted into dishonesty. Institutions that are not seen to punish misbehaviour, not only risk their own credibility, but may encourage students to engage in such behaviour, especially where cheats are seen to benefit from higher grades [7]. Not only does the cheating student miss out on an expected learning experience, but the cheating can also degrade the educational environment, ultimately affecting learning for all students [8].

Perhaps surprisingly, the deterrent effect of detection and punishment remain unclear. Fraser argues that high rates of detection and prosecution deters cheats, and that students are more likely to cheat if they believe cheating is commonplace [19]. An empirical study by Bennett found that punishment was a deterrent to major forms of plagiarism but not necessarily to minor offences [3]. However, Sheard et al. found that fear of consequences did not appear to have any influence on the level of cheating [42].

Beyond honour codes, there are two may strategies that Cotton et al suggest may reduce the incidence of using AI to cheat in assignments:

- Educate students on plagiarism. Educators should explain what is (and what is not) plagiarism and why the behaviour is wrong. Students should be left in no doubt when the use of AI tools are permissible and when it is not allowed. It should also be clear to learners how the use of AI should be cited in their submissions.
- Require students to submit drafts of their work. Assessors can then use tools to detect AI-generated text in early drafts and offer appropriate tuition advice on whether such use is acceptable for the work being undertaken.

6.3 Summary

The intention of our work was not to answer questions regarding assessment design or plagiarism detection; our key focus was on the capabilities of ChatGPT – as an illustrative generative AI tool – to answer current assessment material. Our results help demonstrate the problem, and the need to update assessments, assessment practices, and quality assurance practices. Our analysis demonstrates that, in most cases, across a range of question formats, topics, and study levels, ChatGPT is *at least* capable of producing adequate solutions, and this is of concern for all educators.

7 LIMITATIONS AND FURTHER WORK

We recognise that this study has limitations worthy of further exploration. Firstly, whilst the synthetic solutions often had a similar structure and tone of voice to one another, none of the solutions exactly matched others. The study's small scale means that insufficient synthetic solutions to a single question paper were generated to discover at what point limitations in ChatGPT would eventually produce essentially identical answers. When dealing with hundreds or thousands of synthetic scripts, if there is a saturation point beyond which content is repeated – which is unknowable given the black box nature of ChatGPT – if content is repeated, quality assurance becomes an easier process.

Secondly, the synthetic solutions produced in this experiment represent the laziest of cheats – those who take ChatGPT's outputs and copy and paste it into their assignment answers without attempting to reword the answers or supplement the chatbot's material with their own material. Any such edits or additions are almost certain to increase the difficulty of detecting cheating both through automated processes and the expertise of markers. This is an area ripe for further investigation.

Finally, we only received a set of marks for a relatively small subset of a typical computing curriculum. We have no insights into how well suited ChatGPT and LLMs are for producing robust answers to other areas of the curriculum. This is a pragmatic deficiency and urge others to make use of our methodology to continue contributing to the ongoing research and discussion in this area.

8 CONCLUSIONS AND FURTHER WORK

This study has demonstrated that ChatGPT can achieve passing grades in several computer science undergraduate modules at different levels of study and with differing assessment models. Across a range of undergraduate modules, ChatGPT generated solutions that scored similarly to a random sample of genuine students. ChatGPT performed much less well in the single postgraduate module that was examined with none of the synthetic solutions achieving a passing mark. The study showed that it is not necessary for students to augment ChatGPT responses with their own work to reach at least a passing grade; it is highly likely that students who supplement ChatGPT material would score even higher on their assessment.

Foundational level assessment is extremely vulnerable to ChatGPT. Short-form questions based on simple definitions and applications of knowledge can be answered to a very high level without requiring extensive domain knowledge or skills in query generation. Likewise, questions containing detailed background information; those listing specific points to be addressed in a solution; and those providing templated answers are amenable to high-scoring synthetic solutions. Whilst it is tempting to suggest these issues can be resolved by providing less guidance and requiring longer-form answers; such an approach seriously disadvantages novice students as well as those struggling with module content or language skills.

The LLM challenge to academia strongly resembles that previously created by essay mills. However, the deployment of ChatGPT 'democratises cheating'; it drives the financial cost to zero; greatly reduces the likelihood of being accused of plagiarising content, and; has reduced the time needed to produce solutions to a few minutes. Unlike essay mills, the speed and availability of ChatGPT allows its use in distance examination environments.

In the short-term, a return to face-to-face examinations is a straightforward way of retaining credible assessment; with a longer-term strategy requiring a rethink of the role of assessment as well as how it is conducted, potentially including the greater use of *viva-voce* examinations, live presentations and increased reliance on personal portfolios of work. As a discipline, we are uniquely well placed to assist in both informing institutional policy (as this work has at the OU), and in developing the necessary support material to teach students about the appropriate uses of generative AI. As a community we also need to consider what competencies we expect from our computing students and share best practices as to how to make the best use of generative AI in both our teaching and assessment of those students. Our analysis demonstrates that across a range of question formats, topics, and study levels, ChatGPT is at least capable of producing adequate solutions – but it does provide better answers in some contexts. We need to better explore what assessment types for which topic areas are most resistant to generative AI, to ensure academic quality assurance, while continuing to support students' learning.

ACKNOWLEDGMENTS

We would like to thank John Woodthorpe, Amel Bennaceur, Cecilia Domingo-Merino, Lucas Anastasiou, Nitu Bharati, Venetia Brown, Yaw Buadu, Neil Jones, Jo Buxton, Carole Thornett and Marion Sacharin for sharing their insights. We would also like to thank all of our colleagues who helped support this work, particularly those ALs who were involved in the project. This project was funded entirely by The Open University.

REFERENCES

[1] Željana Bašić, Ana Banovac, Ivana Kružić, and Ivan Jerković. 2023. Better by You, better than Me? ChatGPT-3 as writing assistance in students' essays. DOI:https://doi.org/10.35542/osf.io/n5m7s

[2] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '21), Association for Computing Machinery, New York, NY, USA, 610–623. DOI:https://doi.org/10.1145/3442188.3445922

[3] Roger Bennett. 2005. Factors associated with student plagiarism in a post‐1992 university. *Assessment & Evaluation in Higher Education* 30, 2 (April 2005), 137–162. DOI:https://doi.org/10.1080/0260293042000264244

[4] Laura Bergmans, Nacir Bouali, Marloes Luttikhuis, and Arend Rensink. 2021. On the Efficacy of Online Proctoring using Proctorio. 279–290. Retrieved March 16, 2023 from https://ris.utwente.nl/ws/files/275927505/3e2a9e5b2fad237a3d35f36fa2c5f44552f2.pdf

[5] Emma Bowman. 2022. Scanning students' rooms during remote tests is unconstitutional, judge rules. *NPR*. Retrieved March 17, 2023 from https://www.npr.org/2022/08/25/1119337956/test-proctoring-room-scans-unconstitutional-cleveland-state-university

[6] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (January 2006), 77–101. DOI:https://doi.org/10.1191/1478088706qp063oa

[7] Mark Brimble. 2016. Why Students Cheat: An Exploration of the Motivators of Student Academic Dishonesty in Higher Education. In *Handbook of Academic Integrity*, Tracey Bretag (ed.). Springer Singapore, Singapore, 365–382. DOI:https://doi.org/10.1007/978-981-287-098-8_58

[8] Nathan Brunelle and John R. Hott. 2020. Fix the Course, Not the Student: Positive Approaches to Cultivating Academic Integrity. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education* (SIGCSE '20), Association for Computing Machinery, New York, NY, USA, 1402. DOI:https://doi.org/10.1145/3328778.3372535

[9] Nathan Brunelle and John R. Hott. 2020. Ask Me Anything: Assessing Academic Dishonesty. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education* (SIGCSE '20), Association for Computing Machinery, New York, NY, USA, 1375. DOI:https://doi.org/10.1145/3328778.3372658

[10] Debby R. E. Cotton, Peter A. Cotton, and J. Reuben Shipway. 2023. Chatting and Cheating. Ensuring academic integrity in the era of ChatGPT. DOI:https://doi.org/10.35542/osf.io/mrz8h

[11] Charlie Daly and John Waldron. 2004. Assessing the assessment of programming ability. In *Proceedings of the 35th SIGCSE technical symposium on Computer science education* (SIGCSE '04), Association for Computing Machinery, New York, NY, USA, 210–213. DOI:https://doi.org/10.1145/971300.971375

[12] Paul Denny, Sathiamoorthy Manoharan, Ulrich Speidel, Giovanni Russello, and Angela Chang. 2019. On the fairness of multiple-variant multiple-choice examinations. In *Proceedings of the 50th ACM technical symposium on computer science education*, Minneapolis MN USA, 462–468. Retrieved March 24, 2023 from https://dl.acm.org/doi/abs/10.1145/3287324.3287357

[13] Department for Education. 2022. Essay mills are now illegal - Skills Minister calls on internet service platforms to crack down on advertising - The Education Hub. Retrieved March 22, 2023 from https://educationhub.blog.gov.uk/2022/04/28/essay-mills-are-now-illegal-skills-minister-calls-on-internet-service-providers-to-crack-down-on-advertising/

[14] Martin Dick. 2005. Student interviews as a tool for assessment and learning in a systems analysis and design course. In *Proceedings of the 10th annual SIGCSE conference on Innovation and technology in computer science education* (ITiCSE '05), Association for Computing Machinery, New York, NY, USA, 24–28. DOI:https://doi.org/10.1145/1067445.1067456

[15] Martin Dick, Judy Sheard, Cathy Bareiss, Janet Carter, Donald Joyce, Trevor Harding, and Cary Laxer. 2002. Addressing student cheating: definitions and solutions. *ACM SigCSE Bulletin* 35, 2 (2002), 172–184.

[16] William Dodrill, Doris K. Lidtke, Cynthia Brown, Michael Shamos, Mary Dee Harris Fosberg, and Philip L. Miller. 1981. Plagiarism in computer sciences courses(Panel Discussion). *SIGCSE Bull.* 13, 1 (February 1981), 26–27. DOI:https://doi.org/10.1145/953049.800956

[17] Jarret M Dyer, Heidi C Pettyjohn, and Steve Saladin. 2020. Academic dishonesty and testing: How student beliefs and test settings impact decisions to cheat. *DigitalCommons@COD, College of DuPage* (2020). Retrieved March 20, 2023 from https://dc.cod.edu/cgi/viewcontent.cgi?article=1000&context=testing_pubs

[18] J. Philip East and J. Ben Schafer. 2005. In-person grading: an evaluative experiment. In *Proceedings of the 36th SIGCSE technical symposium on Computer science education* (SIGCSE '05), Association for Computing Machinery, New York, NY, USA, 378–382. DOI:https://doi.org/10.1145/1047344.1047472

[19] Robert Fraser. 2014. Collaboration, collusion and plagiarism in computer science coursework. *Informatics in Education-An International Journal* 13, 2 (2014), 179–195.

[20] Paul Fyfe. 2022. How to cheat on your final paper: Assigning AI for student writing. *AI & SOCIETY* (2022), 1–11.

[21] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, and Noa Nabeshima. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027* (2020).

[22] Aidan Gilson, Conrad Safranek, Thomas Huang, Vimig Socrates, Ling Chi, R. Andrew Taylor, and David Chartash. 2022. How Does ChatGPT Perform on the Medical Licensing Exams? The Implications of Large Language Models for Medical Education and Knowledge Assessment. 2022.12.23.22283901. DOI:https://doi.org/10.1101/2022.12.23.22283901

[23] Will Douglas Heaven. 2020. OpenAI's new language generator GPT-3 is shockingly good—and completely mindless. *MIT Technology Review*. Retrieved March 20, 2023 from https://www.technologyreview.com/2020/07/20/1005454/openai-machine-learning-language-generator-gpt-3-nlp/

[24] Sun Huh. 2023. Are ChatGPT’s knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination?: a descriptive study. *J Educ Eval Health Prof* 20, (2023), 1. DOI:https://doi.org/10.3352/jeehp.2023.20.1

[25] Mohammed Juned Hussein, Javed Yusuf, Arpana Sandhya Deb, Letila Fong, and Som Naidu. 2020. An Evaluation of Online Proctoring Tools. *Open Praxis* 12, 4 (December 2020), 509. DOI:https://doi.org/10.5944/openpraxis.12.4.1113

[26] Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic Detection of Generated Text is Easiest when Humans are Fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, 1808–1822. DOI:https://doi.org/10.18653/v1/2020.acl-main.164

[27] Norman Tiong Seng Lee, Oka Kurniawan, and Kenny Tsu Wei Choo. 2021. Assessing Programming Skills and Knowledge During the COVID-19 Pandemic: An Experience Report. In *Proceedings of the 26th ACM Conference on Innovation and Technology in Computer Science Education V. 1* (ITiCSE ’21), Association for Computing Machinery, New York, NY, USA, 352–358. DOI:https://doi.org/10.1145/3430665.3456323

[28] Lisa Lines. 2016. Ghostwriters guaranteeing grades? The quality of online ghostwriting services available to tertiary students in Australia. *Teaching in Higher Education* 21, 8 (November 2016), 889–914. DOI:https://doi.org/10.1080/13562517.2016.1198759

[29] David J. Malan, Brian Yu, and Doug Lloyd. 2020. Teaching Academic Honesty in CS50. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education* (SIGCSE ’20), Association for Computing Machinery, New York, NY, USA, 282–288. DOI:https://doi.org/10.1145/3328778.3366940

[30] Tony Mason, Ada Gavrilovska, and David A. Joyner. 2019. Collaboration Versus Cheating: Reducing Code Plagiarism in an Online MS Computer Science Program. In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education* (SIGCSE ’19), Association for Computing Machinery, New York, NY, USA, 1004–1010. DOI:https://doi.org/10.1145/3287324.3287443

[31] National Archives. 2022. Skills and Post-16 Education Act 2022. Retrieved March 22, 2023 from https://www.legislation.gov.uk/ukpga/2022/21/contents/enacted

[32] Peter Ohmann. 2019. An assessment of oral exams in introductory cs. 613–619.

[33] Joël Porquet-Lupine, Hiroya Gojo, and Philip Breault. 2022. LupSeat: A Randomized Seating Chart Generator to Prevent Exam Cheating. In *Proceedings of the 53rd ACM Technical Symposium on Computer Science Education V. 2* (SIGCSE 2022), Association for Computing Machinery, New York, NY, USA, 1078. DOI:https://doi.org/10.1145/3478432.3499139

[34] Quality Assurance Agency for Higher Education. 2022. *Contracting to Cheat in Higher Education: How to Address Essay Mills and Contract Cheating (3rd Edition)*. Quality Assurance Agency for Higher Education. Retrieved March 9, 2023 from https://www.qaa.ac.uk/docs/qaa/guidance/contracting-to-cheat-in-higher-education-third-edition.pdf

[35] Quality Assurance Agency for Higher Education. 2023. QAA briefs members on artificial intelligence threat to academic integrity. Retrieved March 29, 2023 from https://www.qaa.ac.uk/news-events/news/qaa-briefs-members-on-artificial-intelligence-threat-to-academic-integrity

[36] Keith Quille, Keith Nolan, Brett A. Becker, and Seán McHugh. 2021. Developing an Open-Book Online Exam for Final Year Students. In *Proceedings of the 26th ACM Conference on Innovation and Technology in Computer Science Education V. 1* (ITiCSE ’21), Association for Computing Machinery, New York, NY, USA, 338–344. DOI:https://doi.org/10.1145/3430665.3456373

[37] Juan Rodriguez, Todd Hay, David Gros, Zain Shamsi, and Ravi Srinivasan. 2022. Cross-Domain Detection of GPT-2-Generated Technical Text. 1213–1233.

[38] Gili Rusak and Lisa Yan. 2021. Unique exams: designing assessments for integrity and fairness. ACM, 1170–1176. Retrieved March 24, 2023 from https://arxiv.org/pdf/2009.01713.pdf

[39] Mihaela Sabin, Karen H. Jin, and Adrienne Smith. 2021. Oral Exams in Shift to Remote Learning. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education* (SIGCSE ’21), Association for Computing Machinery, New York, NY, USA, 666–672. DOI:https://doi.org/10.1145/3408877.3432511

[40] Marian C. Schultz, James T. Schultz, and James Gallogly. 2007. The Management of Testing in Distance Learning Environments. *Journal of College Teaching & Learning* 4, 9 (September 2007), 19–26.

[41] Mike Sharples. 2022. Automated Essay Writing: An AIED Opinion. *Int J Artif Intell Educ* 32, 4 (December 2022), 1119–1126. DOI:https://doi.org/10.1007/s40593-022-00300-7

[42] Judy Sheard, Simon, Matthew Butler, Katrina Falkner, Michael Morgan, and Amali Weerasinghe. 2017. Strategies for Maintaining Academic Integrity in First-Year Computing Courses. In *Proceedings of the 2017 ACM Conference on Innovation and Technology in Computer Science Education* (ITiCSE ’17), Association for Computing Machinery, New York, NY, USA, 244–249. DOI:https://doi.org/10.1145/3059009.3059064

[43] Stability AI. 2023. Stability AI Launches the First of its StableLM Suite of Language Models. *Stability AI*. Retrieved April 25, 2023 from https://stability.ai/blog/stability-ai-launches-the-first-of-its-stablelm-suite-of-language-models

[44] Miriam Sullivan, Andrew Kelly, and Paul McLaughlan. 2023. ChatGPT in higher education: Considerations for academic integrity and student learning. *Journal of Applied Learning and Teaching* 6, 1 (March 2023). DOI:https://doi.org/10.37074/jalt.2023.6.1.17

[45] Wendy Sutherland-Smith and Kevin Dullaghan. 2019. You don’t always get what you pay for: User experiences of engaging with contract cheating sites. *Assessment & Evaluation in Higher Education* 44, 8 (November 2019), 1148–1162. DOI:https://doi.org/10.1080/02602938.2019.1576028

[46] Shea Swauger. 2020. Software that monitors students during tests perpetuates inequality and violates their privacy. *MIT Technology Review*. Retrieved March 17, 2023 from https://www.technologyreview.com/2020/08/07/1006132/software-algorithms-proctoring-online-tests-ai-ethics/

[47] Tamara Tate, Shayan Doroudi, Daniel Ritchie, Ying Xu, and Mark Warschauer Uci. 2023. Educational Research and AI-Generated Writing: Confronting the Coming Tsunami. DOI:https://doi.org/10.35542/osf.io/4mec3

[48] John Warner. 2022. Freaking Out About ChatGPT—Part I | Inside Higher Ed. *Just Visiting*. Retrieved March 7, 2023 from https://www.insidehighered.com/blogs/just-visiting/freaking-out-about-chatgpt%E2%80%94part-i

[49] John Warner. 2023. ChatGPT Both Is and Is Not Like a Calculator | Inside Higher Ed. Retrieved March 7, 2023 from https://www.insidehighered.com/blogs/just-visiting/chatgpt-both-and-not-calculator

[50] Aaron Welborn. 2023. ChatGPT and Fake Citations. *Duke University Libraries Blogs*. Retrieved April 25, 2023 from https://blogs.library.duke.edu/blog/2023/03/09/chatgpt-and-fake-citations/

[51] Will Yeadon, Oto-Obong Inyang, Arin Mizouri, Alex Peach, and Craig Testrow. 2022. The Death of the Short-Form Physics Essay in the Coming AI Revolution. DOI:https://doi.org/10.48550/ARXIV.2212.11661

9 HISTORY DATES

Received May 2023

A APPENDICES

A.1 Question breakdown

Figure 5 shows the overall performance of synthetic solutions in individual questions on TM112, TM129 and TM356. M811 was not included in these diagrams since it is treated as a single long-form question.

TM112 (Synthetic solutions only)

Question

	1	2a	2b	3a	3b	3c	4	5a	5b	6	7
AL1	🟩	🟨	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟨	🟨
Marks	7	19	4	9	5	3	13	7	8	20	5
	Hashing	*Program development*	*Program development*	*Security*	*Security*	*Security*	*Security*	*SQL*	*SQL*	*Social, legal and ethical*	*Module reflection*

TM129 (Synthetic solutions only)

Question

	1a	1b	1c	2a	2b	2c	3a	3b	3c
AL1	🟨	🟩	🟨	⬜	🟨	🟩	🟥	🟧	🟩
AL2	🟨	🟨	🟨	⬜	🟨	🟩	🟥	🟧	🟨
Marks	5	5	5	30	30	30	10	8	7
	Networking	*Operating systems*	*Robotics*	*Networking*	*Operating systems*	*Robotics*	*Module reflection*	*PDP**	*PDP**

TM356 (Synthetic solutions only)

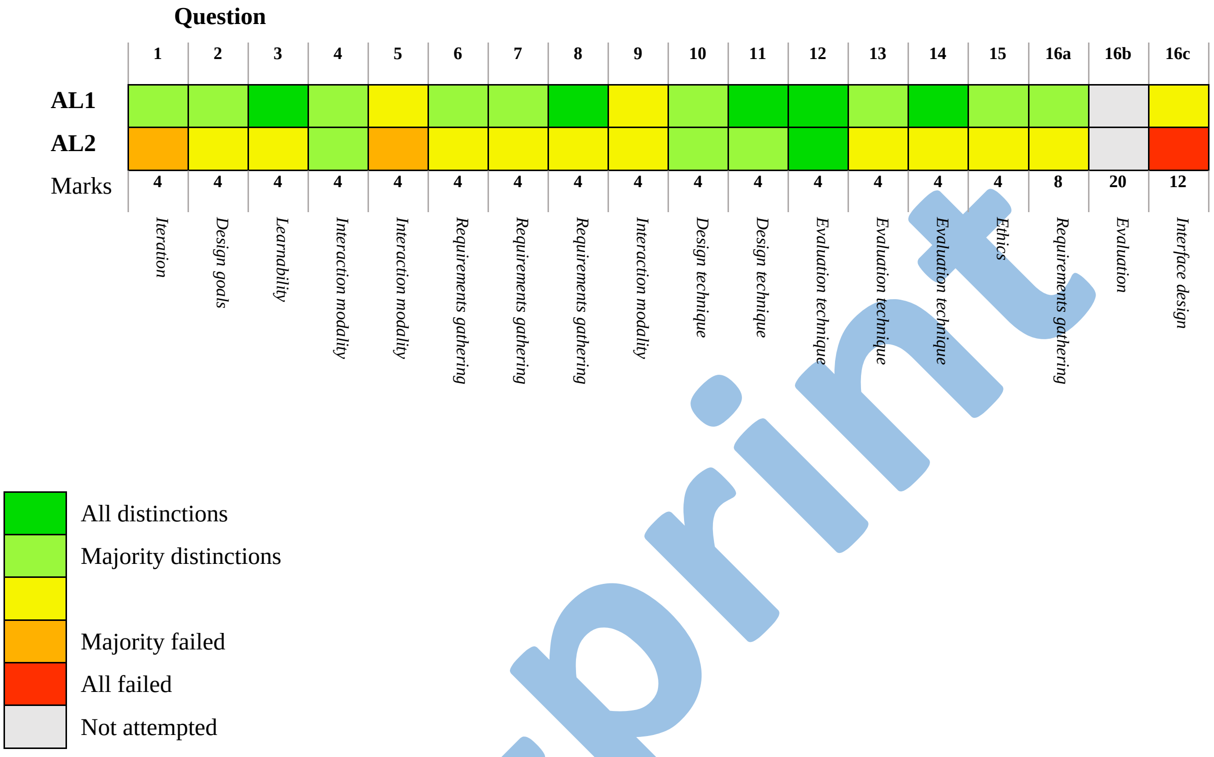

Figure 5: Performance of synthetic solutions in individual questions for TM112 (top), TM129 (middle) and TM356 (bottom).