Gooch, Daniel; Waugh, Kevin; Richards, Mike; Slaymaker, Mark and Woodthorpe, John (2024). Exploring the Profile of University Assessments Flagged as Containing AI-Generated Material. ACM Inroads, 15(2) pp. 39–47.

URL

https://oro.open.ac.uk/97585/

License

(CC-BY-ND 4.0) Creative Commons: Attribution-No Derivative Works 4.0

https://creativecommons.org/licenses/by-nd/4.0/

Policy

This document has been downloaded from Open Research Online, The Open University's repository of research publications. This version is being made available in accordance with Open Research Online policies available from Open Research Online (ORO) Policies

Versions

If this document is identified as the Author Accepted Manuscript it is the version after peer review but before type setting, copy editing or publisher branding
Exploring the Profile of University Assessments Flagged as Containing AI-Generated Material

By Daniel Gooch, Kevin Waugh, Mike Richards, Mark Slaymaker and John Woodthorpe, The Open University

Large language models (LLMs) allow students to generate superficially compelling solutions to assessment questions. Despite many flaws, LLM detection tools provide us with some understanding of the scale of the issue in university assessments. Using the TurnItIn AI detection tool, we present an analysis of 10,725 student assessments submitted in two cohorts during the summers of 2022 and 2023. We observe an increase in the number of scripts flagged as containing AI-generated material. We also present an analysis of the demographic profile of flagged scripts, finding that male students, students with lower prior educational attainment, and younger students are more likely to be flagged.

INTRODUCTION
The abrupt introduction of accessible, freely available LLMs, such as ChatGPT, poses both potential benefits and challenges to CS education. The ITiCSE working group report on Generative AI outlines many benefits, including assistance with code review, the generation of example coding exercises, and alleviating writer’s block [17]. While these benefits are all worth exploring, the immediate future demands we consider the challenges LLMs pose to existing models of academic assessment. Given that compelling answers can be generated using LLMs, existing forms of non-proctored assessment are no longer credible because it will no longer be able to associate a student with certified competencies. As Dick et al. note [5], the inability to identify cheating students can damage the value placed on all academic qualifications and the reputation of associated professions; and potentially endanger society as incompetent graduates enter employment and produce substandard or even dangerous work.

Despite much commentary—both within the popular press and academic circles [20]—there is little empirical work demonstrating current student behaviors with LLMs. In reviewing the literature, the ITiCSE working group evaluated 71 articles, none of which empirically examine the rate of generative AI use within student assessments [17]. The summer of 2023 has been the first opportunity to examine student behavior regarding the end of module assessments following the release of LLMs to the general public.

In this paper we offer an analysis of 10,725 university student assessments, submitted in two cohorts during the summers of 2022 and 2023 respectively. Comparing the data provides empirical insights into the current scale of student use of LLMs. In 2022, generative AI tools were not well known, and much harder to use. By using assessments from 2022 as a baseline, we can distinguish between noise in the detection tool, and changes in student behavior; in essence ensuring that the increase is due to genuine generative AI use, rather than artefacts from the detection tool. We subsequently use demographic data to
Despite much commentary—within the popular press and academic circles—there is little empirical work demonstrating current student behaviors with LLMs.

**BACKGROUND**

Cheating and academic misconduct have been a threat to the integrity of education for as long as there have been examinations. Estimates of scale are difficult to establish but based on a survey of students on an algorithms course, Brunelle and Hott [2] found that rates varied between 29% and 42%, depending on the course topic. Over a range of practices, the main influences on cheating have been found to be time pressure and concerns about failing, while the main countering influences were students’ desire to learn and to know, and demonstrate, what they have learnt [15, 16].

Detection of cheating has previously concentrated on identifying plagiarized material or answers shared between multiple students. Several software products have been developed to automate this detection process and are broadly capable of identifying material plagiarized from online sources as well as evidence of collusion between students in a cohort. However, since LLMs generate novel text each time they are prompted rather than copying it directly from other sources, existing detection mechanisms are no longer sufficient.

Further concerns are raised by remote assessment where students complete work away from the classroom by undertaking exams in their own homes away from the intense supervision found in traditional examination environments. As well as being a staple of established distance education providers, remote assessment was widely instituted as a response to the COVID-19 pandemic and has been retained in some institutions; although concerns over widespread LLM-generated cheating have forced some institutions to revoke such policies [3].

**PERFORMANCE OF LLMs**

There have been some initial investigations into LLMs being able to answer examinations. In computer science, much of the work focusses on software development. Jalil et al. [11] took 31 questions from five chapters of a standard software testing textbook and generated three answers to each question using ChatGPT, also examining whether the answers were better if they were produced in a single thread, or shared context. Their results indicate moderate performance, with 49.4% of the answers deemed correct if generated in a shared context (34.6% for separate threads), and 6.2% partially correct (7.4% for separate threads). Profiling 13 of the 31 incorrect answers, the authors identify a lack of knowledge or making incorrect assumptions as the cause of incorrect solutions.

Researchers at the University of Auckland focused on the capabilities of Codex, a new deep learning model released by OpenAI trained on Python code from GitHub repositories which can take English-language prompts and generate code. Taking 23 questions from two CS1 programming tests, Finnie-Ansley et al. [6] found that Codex could correctly answer all but four of the questions using fewer than 10 responses, with nearly half of the questions (10) being solved successfully on the first attempt. Overall, Codex’s score ranked 17th when compared to scores from 71 students. In a similar study, the authors use Codex to answer more complex CS2 advanced programming questions, and again found the system outperformed students [7].

Outside of an educational setting, code generating LLMs have shown a capacity to generate correct code for common algorithms such as insertion sort and tree traversal [4]. Use studies, however, have shown that the productivity benefits for developers are mixed; making some tasks faster, but in other cases increasing the time required for debugging [23].

Malinka et al. [12] have tested various uses of ChatGPT—as a copy/paste tool, using an interpretation step, and as an assistant—across a range of different assessment types across four security courses. Their findings established that ChatGPT produced answers in Czech that would have passed all courses, except for one case, where simply copying and pasting its outputs led to solutions scoring just below the pass mark (50%) for one module.

We have previously presented a marking exercise where 90 exam scripts (30 generated by ChatGPT, 60 student scripts), from diverse CS modules, were distributed to markers who did not know that some of the scripts were AI-generated [18]. Eighty-eight of the scripts received a pass mark. The authors note that the accessibility of LLMs will significantly impact quality assurance processes as LLMs can generate plausible answers across a range of question formats, topics, and study levels.

**DETECTING LLM-ENABLED CHEATING**

Surveying 171 students, Prather et al. [17] found that 90% of respondents deemed it unethical to auto-generate a solution for a whole assignment as submit it without understanding; dropping to 60% if the student has read and completely understands the generated answer. Only 23% of the respondents deemed it unethical to “auto-generate a solution even for small parts of the assignment.” [17] While changes to assessment practices may limit the risks regarding academic misconduct and generative AI, there needs to be robust measures to prevent students acting improperly.
Alongside the development of LLMs, several developers have released software claiming 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 any given piece of text being the product of an LLM. Unlike conventional plagiarism detection software which can point to identical material elsewhere; these tools cannot demonstrate that the script has not been written by a student.

While there is little published research describing detection of GPT-3/ChatGPT outputs, there is a healthy body of literature of experiments using GPT-2 and other LLMs. Ippolito et al. [10] 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. Rodriguez et al. [19] note: “it is significantly harder to detect small amounts of generated text intermingled amongst real text.” This is an obvious challenge, given that it is likely that most students using LLMs to cheat would use AI 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 built-in to most modern word processors, such as grammar and style checkers 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.

Within computer science education, Orenstrakh et al. [13] have presented an analysis of eight LLM detection tools: GPT2 Output Detector; GLTR; CopyLeaks; GPTZero; AI Text Classifier; Originality; GPTKit; and CheckForAI. Orenstrakh et al. [13] used the eight tools to analyze 164 submissions (124 human-submitted; 30 generated using ChatGPT; 10 generated by ChatGPT and altered using Quillbot). The overall accuracy varied dramatically, with human data being identified between 54.39% and 99.12%; and ChatGPT data between 45% and 95%. False positives ranged between 0 and 52. The authors conclude that “that while detectors manage 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 any given piece of text being the product of an LLM.

RESEARCH GAP
There is currently a lack of empirical analyses on the rate of LLM use in assessments. For the purpose of this work, we have no interest in whether student use of a LLM is considered acceptable or academic misconduct; we are simply investigating the scale of current use. By analyzing 10,725 student records, we provide empirical insight into the current scale of the issue posed by LLMs. An investigation of demographic data provides insight into whether or not current LLM detection tools are highlighting more scripts for students with certain characteristics.

METHOD
Given our research focus, our method is based on analyzing existing data, namely AI detection rates from previously submitted student scripts, and student-reported demographic data. The work was carried out with permission from The Open University human research ethics committee and the data protection team. We are not interested in individual student performance, and our work has had no impact on student academic outcomes or disciplinary records.

TEACHING CONTEXT
The authors all work at The Open University, 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. The modules within our School are designed to be delivered at a 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.

Credit-bearing modules are assessed both during the module through one or more Tutor Marked Assignments (TMAs, broadly equivalent to coursework), and at the end of modules by a written exam or an End of Module Assignment (EMA, similar to coursework but completed as the end point assessment).

For the purposes of this study, we selected seven modules from the undergraduate computing curriculum across introductory, intermediate, and advanced study levels, assessment methods, and topic areas (Table 1). From 2019 onwards, across The Open University, centralized, invigilated examinations have been replaced by at-home examinations. For the purpose of this study, we restricted our work to the final assessment, be that an exam or EMA (except for TM356, where the exam data was unobtainable, leading to us using the final TMA).

LLM DETECTION
The Open University has a policy of scanning all assessment submissions for evidence of academic misconduct using a variety of automated solutions, of which TurnItIn is just one. It has chosen to retain the AI writing detection system offered as part of its existing TurnItIn contract. Given the Open University’s contractual relationship with TurnItIn, it was possible to gather detection data from the tool without raising any further ethical or data protection concerns. As discussed in 2.3 earlier, in the absence of institutional contracts with other providers, legal constraints prevented comparison with other detection systems.

For a given piece of writing, TurnItIn provides an estimated percentage of the material in the writing that has been generated by AI [22]. TurnItIn reports a false-positive (genuine documents incorrectly identified as AI-generated) rate of less than 1%, and a false-negative (AI-generated text that is not recognized) rate of 15%, although little evidence has been released to demonstrate the accuracy of these figures [22]. TurnItIn reports that between April and July 2023, 65 million scripts were processed by the AI tool, with 10.3% of scripts scored as 20% AI-generated content, and 3.3% containing over 80% AI-generated material.

In our previous analysis of 90 exam scripts (30 generated by ChatGPT, 60 student scripts from prior to the release of GPT-3), to consider the accuracy of the TurnItIn detection tools [18], we found that all of the 60 student scripts we scored as containing 8% or less of AI-generated material, with the mean being 0.25%. In comparison, the mean for the ChatGPT-generated scripts was 74.43%, with the lowest script receiving a 28% score. The detection rate was 100%, with a 0% false-positive rate. We conducted our analysis within a week of this study, providing a level of confidence that our results are accurate.

Because of the closed nature of TurnItIn, it is impossible to provide an audit trail for our investigation, as the detection software and parameters can be changed without notice. To ensure that all data for this analysis used the same detection algorithms, all student data was re-uploaded on the same date, and the AI-scores were extracted over a 1-week period.

We must be clear that we are making no academic judgement over whether a high detection score is indicative of academic misconduct; many scripts flagged as containing AI material may have used approved AI tools such as Grammarly to improve writing style; to aid students with conditions such as dyslexia; or those for whom English is not their primary language.

DATA EXTRACTION AND ANALYSIS
At the time of writing (July 2023), TurnItIn does not provide a mechanism for downloading the AI score for a cohort of students; and developing software to extract the data is against their terms of service. As such, the first two authors of this article manually extracted the data. Assessments were re-uploaded to TurnItIn to ensure that a consistent AI-detection algorithm was applied to both cohorts. The authors then copied out the ID for each student assessment into a spreadsheet. The authors opened each script report in TurnItIn and recorded the AI score. TurnItIn sometimes associates a script with an “unknown” flag if it doesn’t meet certain requirements [21]. These scripts were excluded from our analysis.

The Open University data team provided us with the recorded demographic data for students on the selected modules, including age, previous educational attainment, ethnicity, disabilities, gender, and socio-economic status. Our students under-report English as a second language, and therefore the data is not meaningful to use. We note that while some of this data is compulsory as part of student registration, some is optional, and we do not have a complete record — many students choose not to report their personal information.

To investigate the increase in reported AI scores between summer 2022 and 2023, we completed a Chi-square test, given the categorical nature of our data (cohort vs detection label). Chi-square tests were also used for the demographic analysis.

RESULTS
Our first set of results focuses on detection rates within modules. Figure 1 shows the number of student assessments flagged by TurnItIn containing at least 20% or at least 50% of AI-generated material.

Table 2 shows results for module-level Chi-square tests, comparing the expected and actual number of assessments flagged by TurnItIn as containing 20%+ or 50%+ of AI-generated content.

In general, there has been a significant increase in the number of scripts flagged as containing AI-generated material between 2022 and 2023, with all modules showing an increase at the 20%+ level, and most modules showing an increase at the 50%+ level. Overall, the percentage of flagged scripts remains low (a mean of 7% containing 20%+ AI-generated content for 2023 compared to 0% in 2022; and 2% containing 50%+ AI content in 2023 compared to 0% in 2022).
DEMOGRAPHIC RESULTS

For the demographic analysis, the sparse demographics within individual modules would make the results unreliable. Therefore, we looked at the data as a whole, rather than at a module-level.

Table 3 provides an overall summary of the Chi-square tests calculated for each demographic factor. To contextualize these findings, the percentage figures for the number of scripts flagged as containing a specific level of AI-generated material within each demographic category are shown for educational attainment (Figure 2); gender (Figure 3) and age (Figure 4). These are derived from the categories used by The Open University, excluding any categories for which we had no data.

To account for individual students being registered on multiple modules we ran two sets of Chi-square test; one over the full data set (on all Turnitin scores associated with the demographic label), and a second over a reduced data set following the removal of every duplicate student entry.

Table 1: Summary of the modules used in the study

<table>
<thead>
<tr>
<th>Level</th>
<th>Module</th>
<th>Topics</th>
<th>Assessment used in study</th>
<th>Submitted students</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introductory</td>
<td>TM111</td>
<td>Study skills, introductory programming, socio-technical systems, and key networking skills</td>
<td>Final coursework</td>
<td>2005 1685</td>
</tr>
<tr>
<td></td>
<td>TM112</td>
<td>Python programming, computer architectures, and cloud computing</td>
<td>Final coursework</td>
<td>1280 1085</td>
</tr>
<tr>
<td></td>
<td>TM129</td>
<td>Introduction to computer networking, operating systems, and robotics.</td>
<td>EMA</td>
<td>905 822</td>
</tr>
<tr>
<td>Intermediate</td>
<td>TM254</td>
<td>Service management, project management, requirements, and databases.</td>
<td>At-home exam</td>
<td>654 609</td>
</tr>
<tr>
<td></td>
<td>TM351</td>
<td>Accessing, querying, and managing data collections; using data to answer questions about the world, and understanding user needs</td>
<td>EMA</td>
<td>135 184</td>
</tr>
<tr>
<td></td>
<td>TM354</td>
<td>Software development, exploring how software systems are specified, designed, and built from a software engineering perspective</td>
<td>At-home exam</td>
<td>492 480</td>
</tr>
<tr>
<td></td>
<td>TM356</td>
<td>Interaction design lifecycle, covering user needs, requirements, design, and evaluation techniques</td>
<td>Final coursework</td>
<td>191 198</td>
</tr>
</tbody>
</table>

Table 2: Chi-square probability figures, for assessments flagged by Turnitin as containing at least 20% or 50% of AI-generated material for each module in different years.

<table>
<thead>
<tr>
<th>Module</th>
<th>20%+</th>
<th>50%+</th>
</tr>
</thead>
<tbody>
<tr>
<td>TM111</td>
<td>2.00E-20</td>
<td>8.44E-5</td>
</tr>
<tr>
<td>TM112</td>
<td>1.46E-18</td>
<td>4.73E-5</td>
</tr>
<tr>
<td>TM129</td>
<td>4.41E-25</td>
<td>6.98E-14</td>
</tr>
<tr>
<td>TM254</td>
<td>1.71E-10</td>
<td>5.55E-4</td>
</tr>
<tr>
<td>TM351</td>
<td>4.37E-4</td>
<td>0.126</td>
</tr>
<tr>
<td>TM354</td>
<td>7.60E-4</td>
<td>0.153</td>
</tr>
<tr>
<td>TM356</td>
<td>2.87E-3</td>
<td>0.027</td>
</tr>
</tbody>
</table>
Exploring the Profile of University Assessments Flagged as Containing AI-Generated Material

Table 3: Chi-square probability figures, comparing the expected and actual number of assessments flagged by TurnItIn as containing more than 0%, at least 20% or at least 50% of material generated by AI by demographic profile.

<table>
<thead>
<tr>
<th>Demographic factor (group size/non-duplicate group size)</th>
<th>Percentage level of material flagged as containing generative AI material (non-duplicate data)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&gt;0%</td>
</tr>
<tr>
<td>Prior educational attainment (10,593/6,671)</td>
<td>1.76E-5 (3.93E-4)</td>
</tr>
<tr>
<td>Age (10,598/6,676)</td>
<td>9.39E-117 (4.40E-81)</td>
</tr>
<tr>
<td>Gender (10,598/6,676)</td>
<td>8.74E-5 (7.42E-4)</td>
</tr>
<tr>
<td>Disability (10,598/6,676)</td>
<td>0.17 (0.55)</td>
</tr>
</tbody>
</table>

Figure 2: Percentage of student assessments flagged by TurnItIn as containing more than 0%, at least 20% or at least 50% of material generated by AI by prior educational attainment. (Group A: Educational status blank or ‘unknown’; Group B: No formal qualifications or less than two A-levels; Group C: Equivalent of at least two A-levels; Group D: Higher education qualification).

Figure 3: Percentage of student assessments flagged by TurnItIn as containing more than 0%, at least 20% or at least 50% of material generated by AI by gender.
DISCUSSION

Our headline finding is that amongst our student cohort there has been an increase in the number of students flagged as using generative-AI material in their assessments. Overall, the level was lower than anticipated—a mean of 7% for 2023 at 20% compared to 0% in 2022; and 2% for 2023 at 50% compared to 0% in 2022.

The increase in the use of AI generation does not appear to be an artefact of the detection tool in that both data sets were tested with TurnItIn software within a few days of one another. Neither can it be ascribed to an uptake in the use of tools such as Grammarly between cohorts; these were already in use by Open University students prior to the popular use of LLMs. Nor is it likely that changes in text can be ascribed to changes in student writing styles. Instead, the increase in detected text is most likely due to increased student use of generative AI tools, (most likely the highly publicized ChatGPT), in their assessment writing.

It is important to note that we can make no judgement on whether or not this increase in generative AI use in assessment is due to academic misconduct, or students using the tool in an appropriate manner. At the time the samples were submitted for grading, The Open University had not published academic conduct policies regarding appropriate and inappropriate uses for generative AI. Regardless, while academic misconduct remains a concern necessitating a discussion about academic assessments, our data indicates that the current level of adoption of generative AI in assessment is relatively low.

We completely agree with Malinka et al. [12] that improper use of generative AI during a student’s initial learning journey is problematic; many of the skills taught to final-year students require a solid foundation built on previous study, and if that has not been established due to a misuse of generative AI tools, students will not develop these high-level skills. This highlights the need to teach AI-literacy to all students, regardless of their discipline. However, our data (see Figure 1 and Table 2) shows no clear differential uptake by level of study, nor by assessment type or topic area. We acknowledge that despite the significant number of student records analyzed, our sample of modules is simply too small to provide insight into the interwoven nature of level of study, topic of study, and assessment type. As a piece of future work, this would further help institutes shape their assessment practices.

Our demographic data was harder to analyze. Demographic data for any given student will change (all students will age a year between summer 2022 and 2023; disability flags may change as students report them, or diagnoses are made), and the part-time nature of studying with The Open University led to a complex picture, with some students studying at full-time intensity and therefore appearing in multiple modules in both 2022 and 2023. We therefore undertook two analyses: one on the full data set, and a second on a data set with all duplicate students removed. As Table 3 shows, this only affected the significance for one test (>50% AI content - by gender). The overall summary of the Chi-square tests shown in Table 3 is that certain demographic factors—specifically prior educational attainment, age, and gender—show a significant difference between the expected and actual number of assessments flagged by TurnItIn as containing different levels of AI-generated material.

In general, students with lower prior educational attainment have more assessments flagged as containing generative-AI material than those with higher levels of achievement; younger students have more assessments flagged than older students; and male students have more assessments flagged than female students.

Establishing ground truth—whether a student has used a generative AI tool and, if so, for what purpose—is extremely difficult; particularly in the context of a distance university operating at a very large scale. Our inability to directly supervise students at the time they complete assessments renders the

![Figure 4: Percentage of student assessments flagged by TurnItIn as containing more than 0%, at least 20% or at least 50% of material generated by AI by age.](Image)
Our analysis of the demographic data has identified prior educational attainment, age, and gender as factors that lead to significant differences in the number of scripts flagged as containing AI-generated material. We offer this data as a baseline of current performance, one that will be helpful to draw upon as the popularity of LLMs among students increases, assessment strategies are developed, and detection software changes.

Despite the difficulty in interpreting the meaning, it is clear that our data demonstrates a demographic component as to which material is flagged. This requires significant further work to monitor, explore, understand, and mitigate, particularly as the proportion of AI-generated assessment content will continue to increase given the ongoing integration of generative AI-based tools into a wide range of productivity software.

CONCLUSION

Through our analysis of 10,725 student assessments, we provide an empirical analysis on the rate of LLM use in assessments (regardless of whether the use is suitable or academic misconduct). We have demonstrated an increase in detection rates between summer 2022 and 2023 and argue that this is indicative of the informal adoption of generative AI by students to assist in writing their assessment solutions. Having said that, our results show a slightly lower frequency of AI-generated content across cohorts than that reported by TurnItIn itself at the 20% level.

Our analysis of the demographic data has identified prior educational attainment, age, and gender as factors that lead to significant differences in the number of scripts flagged as containing AI-generated material. We offer this data as a baseline of current performance, one that will be helpful to draw upon as the popularity of LLMs among students increases, assessment strategies are developed, and detection software changes.

References


11. Malinka, K., Peresini, M., Finc, A., Hujnák, O., and Janus, F. On the Educational Impact of ChatGPT: Is Artificial Intelligence Ready to Obtain a University Degree?
The #1 Career Destination to Find Computing Jobs.

Check out these new features to help you find your next computing job.

- Access to new and exclusive career resources, articles, job searching tips and tools.
- Gain insights and detailed data on the computing industry, including salary, job outlook, ‘day in the life’ videos, education, and more with our new Career Insights.
- Receive the latest jobs delivered straight to your inbox with new exclusive Job Flash™ emails.
- Get a free resume review from an expert writer listing your strengths, weaknesses, and suggestions to give you the best chance of landing an interview.
- Receive an alert every time a job becomes available that matches your personal profile, skills, interests, and preferred location(s).

Visit https://jobs.acm.org/