A Threat Assessment Framework For Screening the Integrity of University Assessments in the Era of Large Language Models

Alan Hickey, Cathal Ó Faoláin, John Healy, Kevin Nolan, Emer Doheny and Paul Cuffe, *Senior Member, IEEE*

***Abstract*—Since late 2022, the sudden growth in the availability and capabilities of generative artificial intelligence tools, such as Large Language Models, has raised concerns about the threat they pose to the integrity of assessment in educational institutions. Such models are constantly evolving and improving, making the task of understanding exactly *what* they can do more difficult. Recognising this challenge, this paper establishes a Large Language Model exposure framework to qualitatively and quantitatively examine the assessment strategies of university modules to provide a high-level estimated indication of the exposure of these modules to potential dishonest use of such models in the completion of their assessments and coursework. This framework may be used and adapted when planning and reviewing teaching and learning practices and policies.**

I. Introduction

Higher education has dealt with technological advancements in recent decades that have led to a transformation in teaching and learning [1], [2]. Since the arrival of the internet, technology has fundamentally changed how we find and share information - with impacts on both research and teaching [3]. In addition, academic dishonesty has never been easier, while at the same time the wealth of information available has allowed personalised learning on a previously unimagined scale. Large Language Models (LLMs) appear to be the latest escalation in this information revolution. The aim of this paper is to describe a screening framework to proactively identify and quantify the potential for misuse of these models in completing different types of assessments used in university-level engineering coursework.

Plagiarism has always represented a threat to the integrity of academic assessments. However, the growth of online learning combined with the increased abilities of AI models have stoked fears that undetectable plagiarism will become widespread [4], [5]. While artificial intelligence tools for completing homework assignments in mathematics date back as far back as 1964 [6], current LLMs have been shown to be able to solve a wide range of higher-level educational problems [7]–[12] and to evade detection by plagiarism checking tools [13]. This has led to a growth in literature reponding to their capabilities and the consequent implications for teaching and learning in universities [5], [9], [14].

Modern LLMs act as *"text-in, text-out"* tools: often presented as chatbots. They are implemented as neural networks with a transformer architecture, which is based on understandings of human attention [15]. This architecture coupled with vast amounts of training data and supervised learning [16] allows these models to compose cohesive and well-structured text [11], [17], [18]. Such text outputs are composed based on probabilities of word patterns; it could be said that a LLM *pays attention* to the patterns in the data it is trained on to learn how to produce similar textual outputs in the future [15], [16].

While initial work has focused on the powers of these models and the implications this has on education, little literature to date has yet discussed practical ways that higher level educational institutions can assess the potential exposure of their different assessments to plagiarism aided by LLMs. This paper proposes a screening method to loosely gauge the share of coursework that *could* be dishonestly completed by students using LLMs. This screening methodology is intended only to give a high-level overview of possible exposure across a whole portfolio of courses that might be offered by an educational institution. As a qualitative and approximate approach, it is intended only to give a broad summary of which aspects of a portfolio of coursework may be problematic and deserving of further scrutiny.

This paper describes an initial screening framework for assessing the risk of LLM-aided dishonesty across different assessment types. These risk assessments can be aggregated to give numerical ratings of LLM exposure for individual modules or programmes. Understanding the exposure to LLM misuse of a module's assessment structure and the drivers of this should inform a strategic response to LLMs from university leadership.

II. Methodology

A. Survey of LLM Capabilities

To understand *what* LLMs are currently capable of, a review of its abilities in engineering type disciplines was conducted. Their capabilities have proven broad, ranging from the ability to pass a bar exam [10], dermatology and radiology exams [11], and most relevantly, engineering exams [9]. Importantly too, detection tools have so far been unable to consistently identify LLM-generated material [5], [19], and evidence suggests that it may be impossible to reliably distinguish LLM-generated from human-generated original work [17], [19].

The current mathematical capabilities of LLMs are somewhat mixed. Work in [20] holds that while ChatGPT succeeds in completing undergraduate-level mathematical tasks, it performs significantly worse at graduate-level tasks. The authors explain that while the current appraisal techniques of LLM capabilities in mathematics are flawed, updates to LLMs such as GPT-4 have advanced the mathematical competence of LLMs.

A. Hickey, C. Ó Faoláin, J. Healy, E. Doheny and P. Cuffe (paul.cuffe@ucd.ie) are with the School of Electrical and Electronic Engineering, University College Dublin. K. Nolan is with the School of Mechanical and Materials Engineering, University College Dublin. This paper emanated from research funded by Science Foundation Ireland to the SFI Centre for Research Training in Machine Learning (18/CRT/6183) and the Insight Centre for Data Analytics (12/RC/2289 P2) as well the Irish Exchequer, via a grant to UCD by the Higher Education Authority's Strategic Alignment of Teaching and Learning Enhancement scheme.

An in-depth investigation was carried out by seven Australian universities, looking at assessments in courses from mathematics and physics to applied courses in programming, laboratory work, sustainability and renewable energy, amongst others [21]. Their findings conclude that it is possible that students may use LLMs to complete assessments dishonestly where those assessments are not live, in-person or invigilated (e.g., online exams and take-home assignments). They warn that current solutions to mitigate the threat posed by LLM to assessments are based on shortcomings likely to be overcome in the near future - while critical thinking, project-based and in-person assessments will remain beyond the capabilities of LLMs for the foreseeable future [17], [21].

Our literature review revealed the following points:

1) LLMs have a surprisingly good ability in a broader range of disciplines than expected. For example, our own informal experiments found that ChatGPT was able to answer theoretical, applied, general and specific questions about optimisation problems in power system operation and write corresponding code which may be used to solve an economic dispatch problem.
2) The skill set of LLMs is constantly and rapidly improving. What LLMs struggle to do at present, they may succeed at in the not too distant future.

These insights suggested that an in-depth investigation into every assessment task used in every module within a large institution would likely be an inefficient and overly arduous task, yielding the same answer for the majority of assessments; that students could potentially use LLMs to complete these tasks for them, and a potential lack of awareness of the extent of same. Instead, for our screening methodology we decided to take a high-level focus on *where* and *how* these assessments were to be carried out.

B. Analysis of Potential Module Exposure to Irresponsible LLM Use

Reviewing the literature made it clear that it was necessary to examine *where* and *how* assessments within modules were performed in order to develop a high level estimate of exposure to the potential threat to assessment integrity posed by LLMs. LLMs are proven to currently be able to perform engineering-type assignments to a high standard [8], [9], and their current capabilities are only expected to grow [15]–[17]. Therefore the proposed screening methodology sseeks to identify how many coursework assessments are conducted in-person or online. In-person assessments are invigilated and so prevent the ability of LLM-aided plagiarism to go undetected as much as possible: importantly, this fact will remain true with regards to these technologies for the foreseeable future [21].

Based on this, a three-tier classification system was created to assess the proportion of module assessments that take place in different modes. Assessments can be designated as either "*In Person*", "*Blended*", or "*At Home*". *In Person* and *At Home* assessments are self-explanatory, and are deemed to have the lowest and highest potential exposure to LLM misuse, respectively. *Blended* assessments comprise those assessments that require some combination of in-person tasks and at-home write-up (e.g., laboratory sessions or group projects) and fall into the middle exposure category due to the potential for follow-up questions and write-ups to be completed by a LLM.

TABLE I.
ASSESSMENT ACTIVITY LLM EXPOSURE FRAMEWORK

Assessment Type	Assumed Location	Exposure Level	Exposure Score
Attendance	In Person	■	0
Class Test	In Person	■	0
Examination	In Person	■	0
Fieldwork	In Person	■	0
Oral Examination	In Person	■	0
Practical Examination	In Person	■	0
Presentation	In Person	■	0
Seminar	In Person	■	0
Studio Examination	In Person	■	0
Lab Report	Blended	■	2
Multiple Choice Que stionnaire	Blended	■	2
Portfolio	Blended	■	2.5
Group Project	Blended	■	3
Journal	Blended	■	3
Assignment	At Home	■	5
Continuous Assessment	At Home	■	5
Essay	At Home	■	5
Project	At Home	■	5

C. Gathering Screening Data For An Exemplary Instituion

To showcase the high-level screening methodology, this paper will present some general results for University College Dublin's College of Engineering & Architecture. Details of all modules offered by UCD are publicly available on a *module descriptor* website. This website lists the learning outcomes, associated academic staff and assessment structure for each module. Each assessment task is classified as one of 18 types on the descriptor website; these listed in Table I. Using a web scraping script developed in Python, we gathered granular information about each module offered within the College. These descriptors document each module's *assessment strategy*, by which assessment components of eighteen different types are weighted together to give a student's final module grade. The developed Python script mapped these assessment types to an *"Exposure Score"* per the schema we provide in Table I.

The following data was collected from these module descriptors:

- Module Code
- Module Name
- School
- Trimester
- Module Standard Level (0-5)
- Module ECTS[1] Credits
- Each Assessment's Percentage Share of Final Module Grade
- Description of Assessment
- Assessment Type (one of the eighteen categories in Table I)
- Module Enrolment 2022/23[2]

1) Filtering: A number of 'special case' modules existed, wherein the potential exposure to LLM-aided plagiarism was considered low. The two main types of these modules, internship placements and theses, both have constant contact with workplace and academic supervisors throughout the module duration. This greatly reduces the potential of the misuse of LLMs to earn academic credit, and as such were placed in

[1]ECTS, the "European Credit Transfer System" is the system used to assign credits to a programme or module in Ireland, with 1 credit meant to correspond to 20 "learning hours", and 60 credits being the equivalent to a full year of study. 60 Credits is required to proceed to the next year/stage of an undergraduate programme

[2]Module Enrolment for 22/23 was not publicly available, but provided to the university staff and combined with our web scraped dataset.

their own category. Filters were added to remove these from visualisations and reports as desired, to reflect their special circumstances.

2) Recognising Module Importance: Modules which host a large number of students and those which carry a greater amount of ECTS credits are of particular interest when considering the potential exposure to LLM misuse. To capture both the credit-weighting of the module and its assessment components as well as the size of the module, a new metric was devised - Student Credits. Student Credits (SCs) are defined as follows.

$$SCs = Module\ ECTS\ Credits \times Module\ Enrolment \quad (1)$$

SCs allow for modules to be distinguished from each other in terms of student enrolment and ECTS credits. Modules worth the same amount of ECTS credits but with vastly different enrolment figures can easily be differentiated. Likewise, modules with similar enrolment but different ECTS credits can be determined using this metric. The modules offered within the College can be visualised together on a treemap according to their host School and SCs, as seen in Fig. 1.

3) Standard Paths: To examine the variety of assessments encountered by a student over the course of their studies, modules were split into typical pathways, grouped by stage. In UCD, students must take certain "Core" modules and may choose from "Option" and "Elective" modules. Assuming that a typical student will opt for an in-course option or elective module to meet the required 60 ECTS credits per academic year, modules were assigned to example pathways accordingly. This allowed us to model a typical student's experience of university learning, and thus provide LLM-facilitated cheating exposure estimates for their typical module pathways - pathways which were designated "Standard Paths".

4) Calculating LLM Exposure Estimates: While the categorical labels of *At Home*, *In Person* or *Blended* are easy to understand, they are difficult to combine to give an overall view or score to quantify the level of potential exposure of a module and its assessment structure to plagiarism using LLMs. Based on the foregoing arguments about invigilation, an exposure score can be associated with each assessment category in I; these somewhat subjective numerical scores drive the screening analysis, and are listed in the rightmost column.

This numerical rating regime allows for the high level quantification of the level of exposure of a module's assessment structure to cheating by students using LLMs.

An overall module exposure rating can then be calcaulated as the average of individual assessment component ratings, weighted by their share of the final module grade accorded to that assessment, as described in (2).

$$\text{Exposure Estimate} = \sum_{n=1}^{N} R_n S_n \quad (2)$$

R_n is the exposure score for the particular assessment component, as detailed in Table I, S_n is the percentage share contribution of the assessment component to the final module grade and N is the number of assessment components associated with the module.

D. Clustering and Principal Component Analysis

Principal component analysis (PCA) was conducted on the data obtained which describes assessment structures in modules within the College. Specifically, the inputs into this investigation were the weightings of different assessment components for each module, module enrolment numbers for the 2022/23 academic year, module level, module ECTS credits and School. K-means clustering was used to determine the number of clusters that existed within the data. Following this, a plot of the principal components was created, colour coding each of the clusters identified previously.

The aim of this analysis was to ascertain whether the assessment structures of modules give rise to the existence of unique clusters of modules that are assessed in a similar manner. If clusters were based on assessment structures that rely heavily on activities that expose the module to a higher potential for LLM misuse by students, these should be investigated further to safeguard academic integrity.

III. Results

A. Potential Module Exposure to Irresponsible LLM Use

The analysis outlined in Section II facilitates the visualisation of assessment structures within Schools and degree paths at different levels. This visualisation takes the form of a series of charts which detail the breakdown of assessment for a particular set of modules across the three categories corresponding to a School or degree path, for example. An example of such a visualisation showing the breakdown of assessment within one particular degree pathway is shown in Fig. 2. A similar graph for all modules hosted by a particular School of the six within the College is shown in Fig. 3. Similar charts were created for all Schools and degree pathways and circulated internally to relevant members of faculty.

In order to arrive at a quantitative measure of the potential level of exposure for each module within the College to the potential malicious misuse of LLMs, the mapping regime previously described in Equation (2) was used to obtain a numerical score between 0 and 5 for each module, with 0 indicating an estimate of low exposure and 5 indicating an estimate of high exposure. A high-level summary of these LLM-exposure scores is visualised in Figs. 4 and 5.

Figs. 4 and 5 show that while there is generally a reasonable distribution of exposure across all modules, there are a considerable number of modules that have a high potential exposure to LLM misuse for coursework completion. Further investigation revealed that approximately 40% of modules offered by the College are not assessed with an Examination, which is a likely cause of the higher exposure estimates observed. However, analysis of our "Standard Paths", which investigates modules taken by students along an archetypal degree path, showed that these module pathways are much less exposed to LLM misuse than either school or College averages would indicate.

B. Module Exposure Estimate Score Visualisation

Having devised an exposure estimate framework for assessments, a granular exposure estimate rating was determined for each module based on its assessment structure.

This traffic light style analysis is informed by the share of *At Home*, *In Person* and *Blended* assessments accorded to each module and the estimate of potential exposure to LLM misuse for each assessment type described in Table I, and according to the formula shown in the II.

This traffic-light analysis was then combined into a summary for each module which includes details such as its ECTS

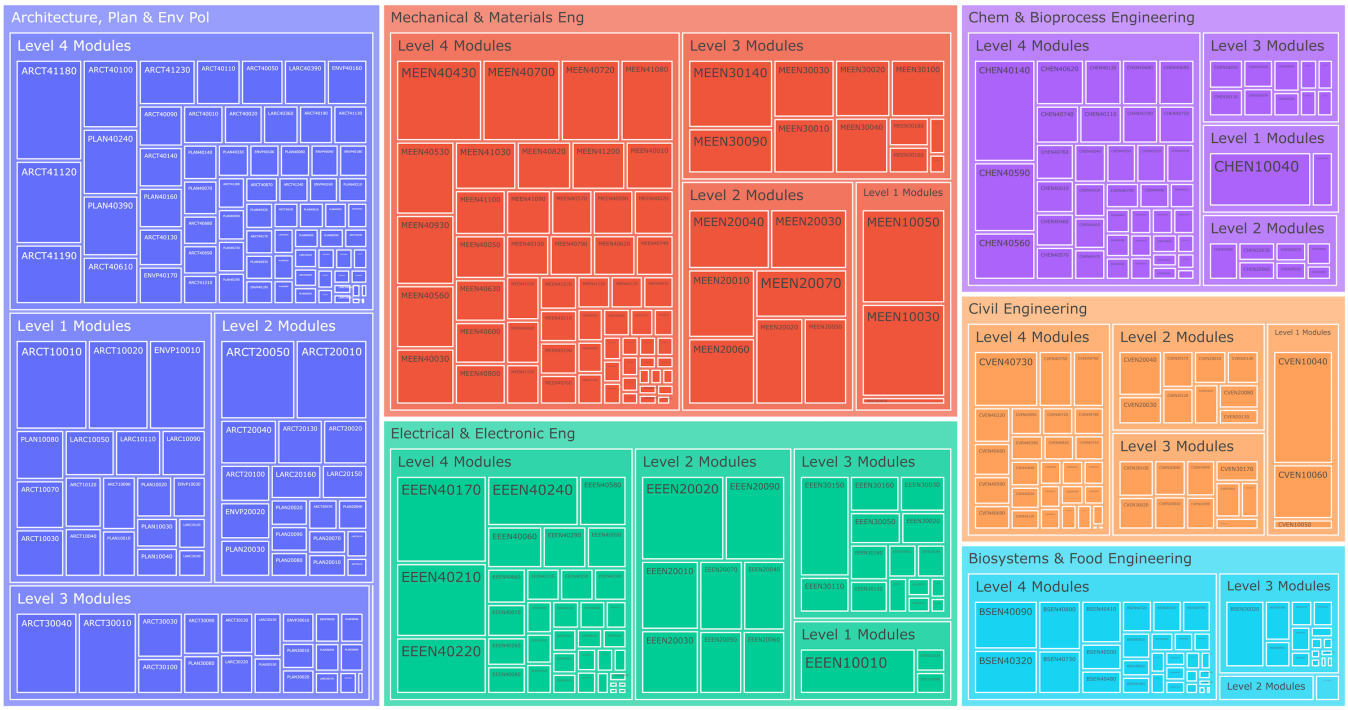

Fig. 1. Treemap showing modules offered within the UCD College of Engineering and Architecture, grouped by School and sized according to Student Credits.

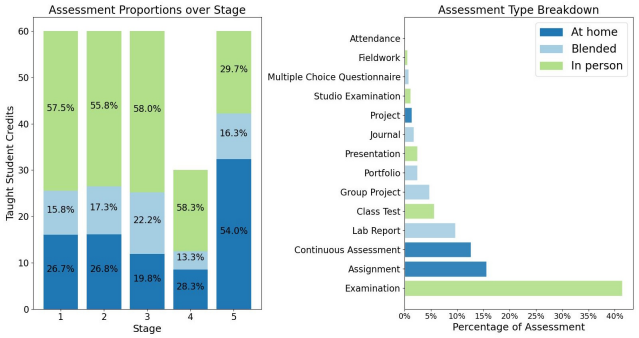

Fig. 2. Assessment types within a particular five year degree pathway offered by the UCD College of Engineering and Architecture. The timeline (on the left) uses the vertical axis to show the portion of a student's annual 60 credits that would stem from each assessment category. The breakdown (on the right) shows the contribution of each assessment type to the total coursework burden completed over the full five years.

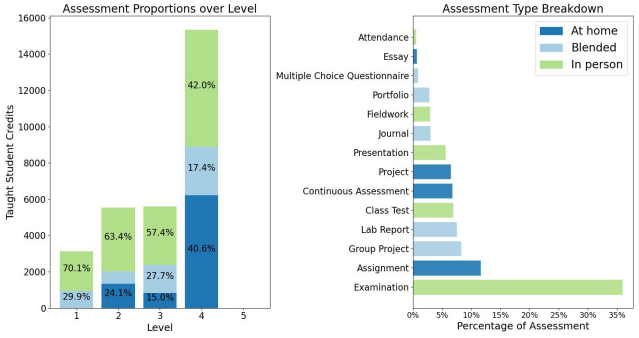

Fig. 3. Breakdown of assessments within modules hosted by one particular School of the six that constitute the College. In the breakdown by level (on the left) the vertical axis is denominated by the Student Credits value, which captures both the enrolment of a module and its ECTS weighting. The barchart (on the right) disaggregates this School's entire coursework portfolio by the different assessment types.

credits, enrolment and assessment structure. A disclaimer and brief explanation of the screening methodology is also included in this summary. These summaries have been circulated to relevant colleagues within UCD.

C. Principal Component Analysis

PCA was conducted on the modules offered within the College in an attempt to identify patterns or "clusters" of modules that that are dominated by a particular set of characteristics. Our aim was to identify module clusters that were assessed in a similar way, and to highlight any that could be deemed to be highly exposed to the potential LLM-aided plagiarism based on their assessment activities.

As can be seen in Fig. 7, PCA did not uncover any particularly interesting trends or clusters within the assessment structures of modules of the College. In fact, Fig. 6 suggests that no especially distinctive cluster groups appear to exist in the data, given the lack of a definite elbow. Based on this, we concluded that we could not find evidence of particular module clusters with high exposure to LLM misuse purely based on any archetypal assessment structure.

It was surprising, though, that none of the clusters were dominated by what the authors considered a "typical" module with a large final exam share and smaller laboratory, assignment and class test components. There are a two potential reasons for this. Firstly, assessment structures may have retained some of the changes introduced in response to enforced remote online learning in 2020 and 2021. Secondly, while all students encounter such "typical" modules, not all students enrol in "niche" modules. Due to the number of inputs to this analysis, it may not have been possible to filter larger, "typical" modules from "niche" modules on the basis of enrolment.

IV. Conclusions

Initial explorations into the capabilities of LLMs revealed that many current academic assessment styles are open to

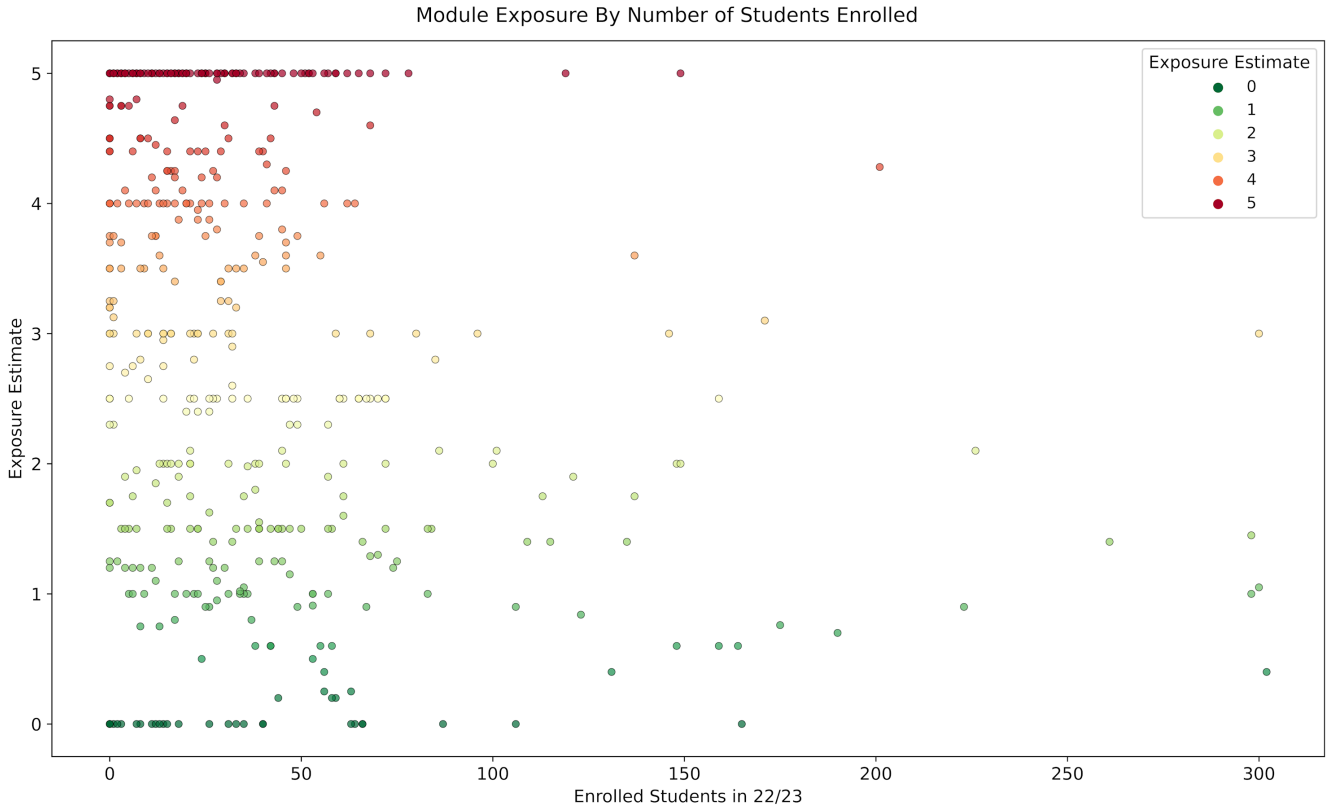

Fig. 4. Scatter plot showing module exposure estimate against number of enrolled students in the 2022/23 academic year for all modules hosted within the College. Each data point represents a module defined by a tuple consisting of the number of enrolled students enrolled in the module and the exposure estimate for the module.

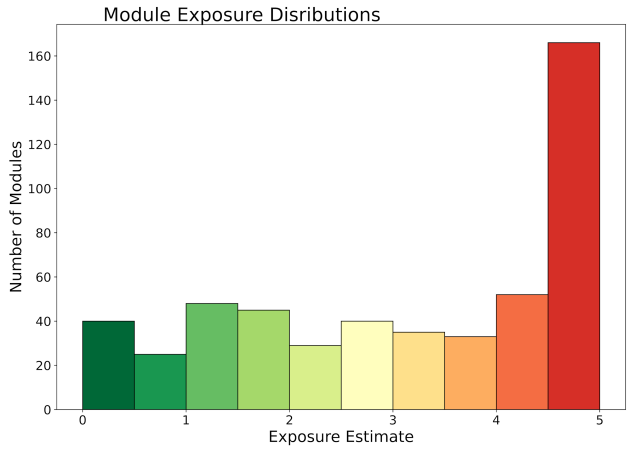

Fig. 5. Histogram showing the distribution of estimated exposure scores across all modules in the College.

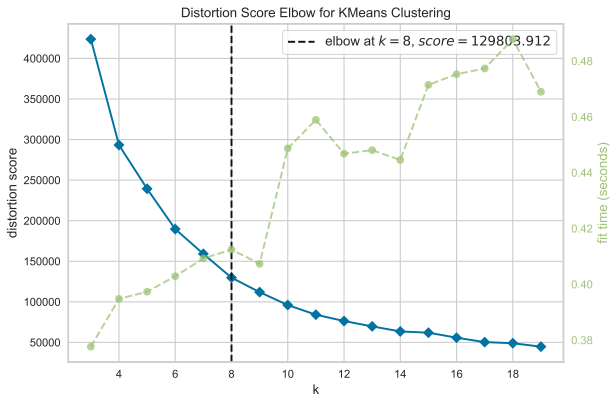

Fig. 6. Graph of distribution score elbow for K-means clustering.

plagiarism from these AI models. While much has been written about how university assessments may need to evolve to mitigate this exposure, there is less awareness around how to attempt to quantify this exposure.

This paper has described a high-level screening framework for gauging the exposure to LLM-facilitated academic dishonesty of a whole portfolio of university coursework.

First, module assessment information was obtained from the publicly available module descriptor website. This was supplemented by module enrolment numbers. By classifying assessment types as either *In Person*, *Blended*, or *At Home*, a numerical and visual exposure estimate for each module were determined.

Use of a derived measure, "student credits", as well as identifying standard student stream paths allowed us to quantify important modules, while filters were created that could remove modules of types that were deemed less susceptible to LLM-facilitated cheating such as internships and theses. These tools allowed the analysis to take place at different scales, giving both a granular overview of every individual module, as well as larger analyses on stream, school and college levels.

On one hand, the analysis showed that the modal exposure

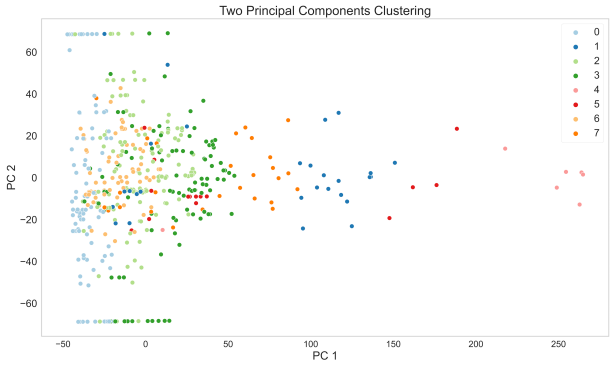

Fig. 7. Scatter plot of the first two principal components showing the eight clusters, illustrating the lack of distinct grouping patterns across modules.

estimate interval was the highest category (4.5 - 5). This is mainly driven by a significant number of modules within the College that are not assessed by an Examination assessment component. However, taking a student's perspective along standard degree pathways shows that the potential exposure to LLM misuse for modules encountered by students is less. As students specialise along these paths, the number of modules on offer grows considerably. These smaller modules place greater make use of different assessment methods aside from *In Person* assessments. This contributes to the apparent higher exposure to modules in the College as a whole.

This work should not be considered as merely a suggestion to replace higher risk assessments such as essays, assignments and projects with strictly *In Person* assessments such as exams as this would not be of benefit to the students concerned. Likewise, it is important to state that the exposure estimates calculated do not endorse any particular assessment regime, nor do they guarantee that a module with a low exposure estimate score of 0 is totally "safe" from potential LLM misuse by students completing assessments. Equally, an exposure estimate score of 5 does not imply criticism or lack of suitable assessment difficulty. Balancing the risks of a rich assessment structure, student honesty and plagiarism are vital in equipping students to make full use of acquired skills.

References

[1] D. Akaslan and E. L.-C. Law, "Measuring teachers' readiness for e-learning in higher education institutions associated with the subject of electricity in Turkey," in *2011 IEEE Global Engineering Education Conference (EDUCON)*, 2011, pp. 481–490. DOI: 10.1109/EDUCON.2011.5773180.

[2] D. von Grünigen, F. B. de Azevedo e Souza, B. Pradarelli, A. Magid, and M. Cieliebak, "Best practices in e-assessments with a special focus on cheating prevention," in *2018 IEEE Global Engineering Education Conference (EDUCON)*, 2018, pp. 893–899. DOI: 10.1109/EDUCON.2018.8363325.

[3] J. Rudolph, S. Tan, and S. Tan, "ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?" *Journal of Applied Learning and Teaching*, vol. 6, no. 1, 2023.

[4] M. Khalil and E. Er, "Will ChatGPT get you caught? Rethinking of plagiarism detection," *arXiv preprint arXiv:2302.04335*, 2023.

[5] M. Halaweh, "ChatGPT in education: Strategies for responsible implementation," 2023.

[6] D. Boblow, "Natural language input for a computer problem-solving system," *Semantic Information Processing*, pp. 146–226, 1968.

[7] B. Qureshi, "Exploring the use of ChatGPT as a tool for learning and assessment in undergraduate computer science curriculum: Opportunities and challenges," *arXiv preprint arXiv:2304.11214*, 2023.

[8] J. Blocklove, S. Garg, R. Karri, and H. Pearce, "Chip-chat: Challenges and opportunities in conversational hardware design," *arXiv preprint arXiv:2305.13243*, 2023.

[9] V. Pursnani, Y. Sermet, and I. Demir, "Performance of ChatGPT on the US Fundamentals of Engineering exam: Comprehensive assessment of proficiency and potential implications for professional environmental engineering practice," *arXiv preprint arXiv:2304.12198*, 2023.

[10] D. M. Katz, M. J. Bommarito, S. Gao, and P. Arredondo, "GPT-4 passes the bar exam," 2023. DOI: 10.2139/ssrn.4389233.

[11] E. Kasneci *et al.*, "ChatGPT for good? On opportunities and challenges of large language models for education," *Learning and Individual Differences*, vol. 103, p. 102 274, 2023.

[12] R. Bhayana, S. Krishna, and R. R. Bleakney, "Performance of ChatGPT on a Radiology Board-style Examination: Insights into Current Strengths and Limitations," *Radiology*, vol. 307, no. 5, e230582, 2023. DOI: 10.1148/radiol.230582.

[13] D. O. Eke, "ChatGPT and the rise of generative AI: Threat to academic integrity?" *Journal of Responsible Technology*, vol. 13, p. 100 060, 2023.

[14] X. Zhai, *ChatGPT user experience: Implications for education*, 2022.

[15] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon *et al.*, Eds., vol. 30, Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[16] N. Stiennon *et al.*, "Learning to summarize with human feedback," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., 2020, pp. 3008–3021. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/1f89885d556929e98d3ef9b86448f951-Paper.pdf.

[17] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 610–623.

[18] B. Qureshi, "Exploring the use of ChatGPT as a tool for learning and assessment in undergraduate computer science curriculum: Opportunities and challenges," *arXiv preprint arXiv:2304.11214*, 2023.

[19] M. Khalil and E. Er, "Will ChatGPT get you caught? rethinking of plagiarism detection," *arXiv preprint arXiv:2302.04335*, 2023.

[20] S. Frieder *et al.*, "Mathematical capabilities of ChatGPT," *arXiv preprint arXiv:2301.13867*, 2023.

[21] S. Nikolic *et al.*, "ChatGPT versus engineering education assessment: A multidisciplinary and multi-institutional benchmarking and analysis of this generative artificial intelligence tool to investigate assessment integrity," *European Journal of Engineering Education*, pp. 1–56, 2023.