‘Frenemy’ of progress? Investigation of the disruptive impacts of generative pre-trained transformers (GPT) on learning and assessment in higher education

KOLADE, Seun <http://orcid.org/0000-0002-1125-1900>, OWOSENI, Adebowale and EGBETOKUN, Abiodun

‘Frenemy’ of progress? Investigation of the disruptive impacts of generative pre-trained transformers (GPT) on learning and assessment in higher education

Oluwaseun Kolade[1], Adebowale Owoseni[2] & Abiodun Egbetokun[3]

Introduction

The application of artificial intelligence (AI) in education has been a subject of growing interest over the past decade. This is especially the case in language learning, where AI agents have been deployed to provide bespoke instructions for students in large classes and offer targeted and unlimited practice opportunities that are otherwise unrealisable in traditional classroom settings (Wang *et al.*, 2023). Chatbots are demonstrably effective as pedagogical tools, offering language learners particular advantages as writing partners, in terms of the variety of language they bring to the writing process, the prompt feedbacks for students, and the stress-free interactions with students in the face of inevitable mistakes (Guo, Wang and Chu, 2022).

In recent years, rapid advances in artificial intelligence have led to the emergence of generative pre-trained transformer 3 (GPT-3), a state-of-the art autoregressive language models which offerings and capabilities far supersedes previous models of chatbots. With 175 billion parameters at its command, GPT-3 is one of the largest and most powerful language processing AI models available (Dale, 2021). With its vast and versatile capabilities, GPT-3 has been used to produce academic essays, technical reports, comedy scripts and poetry, to mention a few. GPT-3 power lie in its unprecedented capabilities to mimic human produced texts.

As the enormous capabilities of GPT-3 capture public imagination and fever-pitch interest, it is also beginning to focus the minds of stakeholders on its implications, consequences and potential dark sides. In this paper, we focus attention on the implications of the new technology for learning and assessment in the higher education sector. Over the past decades, universities have grappled with the challenge of essay mills, a problem that has been significantly exacerbated by advertent and ubiquity of the internet (Crook and Nixon, 2021). Given that academic essays are a mainstay of assessments in colleges and universities, the problem of essay mills has become intractable, even in the wake of web-based plagiarism detection systems such as Turnitin.

GPT-3 is a disruptive game changer that further exacerbates the intractable challenge of essay milling, but also potentially offer new and promising pathways to learning and assessment. First, the emergence of GPT-3 appears to have “democratised cheating”, as students are now able to generate original essays in seconds and at little or no cost, and without recourse to essay mills. Conversely, revolutionary advances in AI invariably push the frontiers of learning in the age of digital transformation, offering new opportunities to rethink and deepen learning and assessment in higher education.

Given the above, we raise two related and sequential research questions in this paper. Firstly, we ask: what is the impact of GPT-3 on the evaluation of students’ learning? Secondly, what new opportunities are offered by GPT-3 to enrich students’ learning experience? The first question is empirical, the second conceptual. The empirical component focuses on evaluation

[1] Corresponding author: Professor Seun Kolade, Sheffield Business School, Sheffield Hallam University. Email: seunkolade2014@gmail.com
[2] School of Computer Science and Informatics, De Montfort University
[3] Leicester Castle Business School, De Montfort University

of students' learning, rather than actual learning, which is outside the scope of our research design and data. Following on this, the conceptual component focuses on potentials and opportunities of GPT-3 for students learning in the context of the new knowledge economy.

The rest of the paper is organised as follows. First, we present a review of the extant literature on learning and assessment in higher education, the use of AI in education, and pathways to new forms of learning and assessment. This is followed by a description of the study's methodology, including an overview of data collection using Chat GPT, and analytical procedures using Turnitin, among others. Next, we present the results and offer empirical explications and conceptual insights in the light of the data. Finally, we conclude the paper with an overview of key findings, practical implications for pedagogy, and recommendations for future studies.

Literature review

Artificial intelligence, learning and assessment in HE

Historically, assessments have been used in college and universities evaluate and certify students' learning (Rawlusyk, 2018).Thus, the two main purposes of assessment have been identified as: facilitation of learning on the one hand, and certification of achievement, on the other. These can be achieved through formative and summative assessments. Formative assessment is defined as an ongoing process of continuous exchange of information and feedbacks between learners and teachers with the aim of modifying teaching practice and learning activities to help students learn more effectively (Dixson and Worrell, 2016). In other words, in formative assessments, learning activities and outcomes are continually shaped (formed) through a dialogic, interactive process in which learners are actively co-opted to identify what is working, what needs to be improved, and how it can be improved for better learning experience of students. Summative assessments, on the other hand, are typically used to evaluate students' learning at the end of a unit of learning (Goss, 2022). As such, they are typically teaching centred and used to establish learners' academic progress based on some established criteria (Dunn and Mulvenon, 2009).

In the wake of Covid-19 pandemic, remote and asynchronous teaching and learning have grown popular in higher education (Lockee, 2021). These modes of delivery underlie the digital transformation that is now taking place in education (Gallagher and Palmer, 2020). Advances in the broad field of computing, more specifically in artificial intelligence, have led to the development of tools that possess unprecedented transformative potential. For instance, the new GPT-3 can generate curriculum content, fix bugs in computer codes and write complex passages that compare to human output (Sharples, 2022).

Prior to GPT-3, purpose-built AI tools have been successfully applied for assessment in the context of research and education (Lagakis and Demetriadis, 2021). For instance, Checco et al. (2021) describe an experiment in which an AI system accurately predicts the review outcome of the peer review process. Among a set of students learning English as a foreign language in China, (Wang, 2022) reported that an AI-enabled system performed better than human teachers in terms of feedback effectiveness and impact on students' learning ability. Similarly, McNamara, Crossley and Roscoe (2013) describe a sophisticated AI-enabled tool that seems to match the performance of human instructors in providing essay writing tutorship.

Despite the ongoing digital transformation and the opportunities offered by AI, assessments in higher education remains fundamentally unchanged. This is probably due to the psychological reaction premised on the tendency that "most people like things to be comfortable and familiar" (Craine, 2007:44). Today's assessment tests in higher education follow a tradition that started in the 19th century (Kruse, 2006) which relies on written essays and reports as the gold standard of student assessment in tertiary education. However, written essays are alleged to have limited objectivity and high susceptibility to cheating through outright plagiarism or using paid writers (Newton, 2018).

The written essay is arguably the most widely used form of summative assessment in higher education today, although it has been the subject of ongoing debate. Relative to sit-in exams, essays are thought to have a tendency to better elicit aspects of applied learning, especially critical thinking and appreciation of how abstract concepts are related(Covic and Jones, 2008) Essays are also known to reduce students' pressure for rote learning in contrast to traditional examinations which often promote rote learning (Chuderski, 2016). Indeed, higher education students prefer assessments that build on their skill set, gives them some power of choice and allows for creativity. Well-designed essay tasks are believed to exhibit these attributes (Lynam and Cachia, 2018). Perhaps for this reason, AI-enabled tools such as OpenEssayist (Whitelock et al., 2013) and Writing Pal (McNamara, Crossley and Roscoe, 2013) have been developed specifically to support students in writing essays.

However, essays are inherently weak, ultimately. Because of the risk of collusion and copying, essays may undermine effective learning. They are also particularly prone to unethical practices such as plagiarism and wholesale ghost writing (Newton, 2018; Sharples, 2022). The advent of transformer AI such ChatGPT which can generate highly original text at very little cost in terms of time and funds, introduces a new set of challenges that existing tools and models may not be well equipped to deal with. We argue in this paper that these new AI tools can transform both learning and assessment in higher education, and therefore require paradigmatic shifts in current models in order to make the best use of them while also anticipating and mitigating any risks that they bring.

Potential implications of AI on learning and assessment practices

In the face of rising AI possibilities, it remains to be seen how teaching and assessment models will evolve. GPT-3 in particular holds tremendous potentials for positive and negative consequences. Not only does it possess a remarkable ability to generate human-like responses, it can also produce complete, intelligible and logical essays. A positive consequence of this ability is that it can help provide useful starting ideas for written work. On the negative side, since there are no known tools to reliably distinguish AI-generated text from human-generated text, students may get away with passing off AI-generated text as their own work (Sharples, 2022). In a world where cheating on written essays is already high (Newton, 2018) this poses an even greater challenge.

The pedagogical potential of AI tools like GPT-3 is also receiving attention. A meta-analysis by Bibauw et al. (2022) indicates that AI chatbots are strikingly effective for language learning. In a systematic review of 74 publications on chatbots' application in education, Wollny et al. (2021) highlighted several benefits of AI chatbots. These include, among others, scalability and accessibility. The advantages notwithstanding, some downsides of AI use in education are already recognised. For instance, a real ethical and practical problem arises from the possibility

that learners and teachers alike may bypass genuine knowledge exchange if they rely excessively on AI tools for content generation. AI tools are also unable to replicate certain 'human' components of pedagogy, such as empathy, mindfulness and helpfulness (Tack and Piech, 2022). Moreover, algorithmic bias may limit the objectivity of AI tools, a limitation that users are unlikely to be aware of (Checco et al., 2021).

In the case of ChatGPT-3, it is a cutting-edge AI transformer model known as a Large Language Model (LLM), with 175 billion parameters at its disposal, trained with large amounts of data to understand and process language in similar ways that humans do. This includes the ability to engage in discussions, dialogue, frame ideas, and communicate effectively. ChatGPT-3 was specifically trained with over 570 gigabytes of data (Tamkin et al., 2021), moreover, the self-supervised learning approach used in the development of ChatGPT-3 allowed it to improve its capabilities and perform tasks such as programming, mathematical computations, and language translations with a few or no specific training examples (Brown et al., 2020). The model can self-learn from large amounts of unlabelled data. by absorbing large volumes of text and predicting missing words and sentences. OpenAI leveraged one of the top five supercomputers in the world to train ChatGPT-3, using a specially made computer having more than 285,000 CPU cores, 10,000 GPUs, and 400 gigabits per second of network connectivity across the GPUs (Langston, 2021).

The deployment of transformer AI systems such as ChatGPT for academic essay writing and other forms of content generation has renewed otherwise longstanding conversation about the place of knowledge in the pyramid of learning outcomes. In a framework of assessment originally proposed for clinical training, knowledge (know what) is at the base of the pyramid, identified, in effect, as the starting point of assessment. Know what is followed by know how, or competence, and this in turn is progressively followed by performance (show how) and action at the top of the pyramid (Miller, 1990). Thus, in this paper, we note that GPT-3 is being increasingly used by students to generate knowledge and thereby achieve the "know what" outcome at the base of the pyramid, in a process that is difficult to stop or track. Given this, there is a case to be made for full acceptance and even active support for the use of these transformer AI systems, while restructuring assessments to focus on the "know how" (competence) performance levels of assessment. In these higher levels, the focus of assessment is retuned from evaluation of learners' knowledge to appraisal of what they can do with the knowledge. This approach will, of course, raise new operational questions about how this can be achieved in practice.

Methodology

A quasi-experiment approach was employed to investigate the impact of GPT-3 on student evaluation and the new opportunities it offers for enhancing students' learning experience. The experiment involved five persons, consisting of three researchers and two other participants. All participants opened a ChatGPT account two weeks prior to the experiment and were tagged as Account 1 to 5, respectively*:* Account 1 instructed ChatGPT to "*write an essay on the digital transformation of the health sectors in the global south. It will be useful to provide suggestions on how to improve digitally-enables healthcare delivery*". This resulted in the generation of Essay 1. This process was repeated five times, generating Essays 2 to 6 from the same user account 1. The same instruction was repeated from four other unique accounts. When ChatGPT was instructed to generate citations, it included a number of in-text references. When pressed

further to generate a full reference list, it apologised, saying in effect that the in-text citations it previously generated did not, in fact, exist. The five distinct essays were independently graded by two academics recruited for this purpose. They graded the essays as normal student essays, being unaware of the ChatGPT experiment.

Results and discussion

Results

Table 1 contains the results of the plagiarism evaluation of the outputs from Stages 1 and 2 as well as the human assessment undertaken in Stage 3. The Turnitin similarity index of the essays ranged from 4% to 99%. A low index was observed for the first essays generated for each user account. For Account 1, which generated 6 essays, the similarity index increased significantly after the first essay, from 4% to 86% and then to 99%. Human evaluation of the essays produced an average score of 75.6% for Essays 1, 7, 8, 9 and 10. With a score of around 80% each, Essays 1, 7 and 8 were adjudged by the human assessor to be of comparable quality. Essays 9 and 10 scored lower mainly because they were adjudged to be weaker in the use of theories and concepts.

Table 1: Analysis of user accounts and similarity Indexes of essays

S/N	Submission	Sequence	User Account	Turnitin Similarity %	Human Evaluation %	Feedback summary
1	Essay 1	1st	User 1	4%	82%	Overall structure and presentation (20/25), theory and concepts (21/25), Coherence (20/25) and Conciseness (21/25)
2	Essay 2	2nd		86%	Not Applicable	Not evaluated by Human
3	Essay 3	3rd		99%		
4	Essay 4	4th		99%		
5	Essay 5	5th		88%		
6	Essay 6	6th		97%		
7	Essay 7	7th	User 2	18%	80%	Overall structure and presentation (20/25), theory and concepts (18/25), Coherence (21/25) and Conciseness (21/25)
8	Essay 8	8th	User 3	19%	80%	Overall structure and presentation (19/25),

						theory and concepts (21/25), Coherence (22/25) and Conciseness (18/25)
9	Essay 9	9th	User 4	24%	66%	Overall structure and presentation (20/25), theory and concepts (12/25), Coherence (18/25) and Conciseness (16/25)
10	Essay 10	10th	User 5	17%	70%	Overall structure and presentation (21/25), theory and concepts (13/25), Coherence (18/25) and Conciseness (18/25)

Despite coming from unique user accounts and showing low Turnitin similarity, Essays 1, 7 and 8 were adjudged by the human assessors to be similar in content albeit with modest differences. This is to be expected since Turnitin focuses on textual similarity while a human assessor would naturally focus on similarity of substantive content. The following excerpt from the feedback illustrates this point:

> *The papers are highly similar with each students identifying related challenges and solutions. The differences lie slightly in the manner of presentation of the harnessed solutions. Student 7 was specific in mentioning the role of Govt and private operators (other stakeholders) in revolutionising digital healthcare delivery; student 8 used "we" in harnessing digital health solutions in global south while student 1 mentioned significantly "the government" in improving the digital transformation of the health sector.* - ***Feedback on Essays 1, 7 and 8***

The feedback on Essays 9 and 10 provides more specific feedback on the strengths and weaknesses of the essays. It shows, for instance, that the essays had a good structure, relevant content but lacked conceptual grounding or proper references.

> *Overall, the essay has a structure, with a brief introduction of the topical areas, a body presenting the different varieties of digital transformation to be adopted as well as a conclusion summing up the points and summarising the argument advanced in helping the global south. Although to some extent the concepts of digital transformation well addressed there is no reference to what enables digital transformation, such as technology diffusion, or even absorptive capacity. Ideally, the essay should have drawn on theoretical concepts to explain how digital transformation would have worked in a deprived rural community in the global south as its assuming this is possible with all*

the limited resources, infrastructure etc. that is essential for digital transformation. Generally, the essay has a logical flow with some transitions from the different paragraphs, however, can be improved. Additionally, a good attempt is made in presenting the arguments with conciseness. However, the attempt to present three digital transformations prevent the essay from addressing in depth how these technologies can be implemented given the challenges in the global south - ***Feedbacks on Essays 9 and 10***

Next, we document the instructions given to ChatGPT at Stage 4 of the experiment, as well as our observations (Table 2). It comes out from the observations that ChatGPT seems to struggle with referencing, a point that the second human assessor also picked out. Figure 2 summarises the findings of the thematic analysis of Stage 4 output. The analysis reveals 12 clear themes across the five article genres. Other characteristics such as word count, citation, structure, and tone were also recorded. We observe that the outputs are of varying length, the essay (593 words) being the longest and the poem (261 words) being the shortest. We did not issue any word count criteria so it is safe to assume that ChatGPT decided for itself. We also observe that all the outputs used an appropriate structure and tone.

Table 2: Overview of Chat GPT outputs

Tasks	Instruction	Key Observation
1	Write an **essay** on "**the digital transformation of the health sectors in the global south**". It will be useful to provide suggestions on how to improve digitally enabled healthcare delivery.	ChatGPT did not finish the essay all at once; the researcher had to prompt the bot to finish it. There was no citation, but a reference was provided: *Kroneman, M., Brouwer, W., & Scott, A. (2016). Digital health in low- and middle-income countries: from promises to impact. Lancet, 387(10022), 1040-1050.*
2	Write a **report** on "**the digital transformation of the health sectors in the global south**". It will be useful to provide suggestions on how to improve digitally enabled healthcare delivery and **cite relevant literature**.	ChatGPT's citations include (Kroneman et al., 2016), (Raza et al., 2018), and (Takahashi, et. al., 2019). When ChatGPT was asked to provide references for the citations, it replied, "*I apologise, as I made an error in my previous response. (Kroneman et al., 2016) and (Takahashi et al., 2019) are references that do not exist in academic literature. In any academic report, it is critical to use credible sources. JSTOR, Pubmed, Scholar Google, and other databases can be used to find credible and recent articles, journals, and papers relevant to your topic and use them as references*."
3	Write a **blog post** on "**the digital transformation of the health sectors in the global south**". It will be useful to provide suggestions on how to improve digitally enabled healthcare delivery.	The essay was finished all at once.
4	Write an **editorial message** for me on "**the digital transformation of the health sectors in the global south**". It will be useful to provide suggestions on how to improve digitally enabled healthcare delivery. The	In comparison to other genres, the message was more focused on the problems.

	editorial is for **a local newspaper**.	
5	*Please write **poetry** for me on "**the digital transformation of the health sectors in the global south**". The **poetry will be presented at the WHO conference***	ChatGPT stopped writing the poem after the sixth verse, however, the last verse did not represent a conclusive chapter, as a result, the researcher gave ChatGPT a prompt, and it wrote four more verses. The poem's tone is positive and motivating, inspiring positive actions.

Title: The digital transformation of the health sectors in the global south						
Theme		**Genre**				
		Essay	Report	Blog Post	Editorial	Poetry
1	Digital transformation awareness and urgency	✓	✓	✓	✓	✓
2	Itemised challenges of healthcare delivery	✓	✓	✓	✓	✓
3	Itemised positive impacts of digital transformation	✓	✓	✓		
4	Use of Electronic Health Records (HER)	✓	✓	✓		✓
5	Use of Telemedicine	✓		✓		✓
6	Lack of Infrastructure	✓	✓	✓	✓	✓
7	Shortage of trained personnel	✓	✓	✓	✓	
8	Digital literacy	✓	✓		✓	✓
9	Lack of standardization and interoperability	✓	✓	✓	✓	✓
10	Need for increased investment by government	✓	✓	✓	✓	✓
11	Need for government to prioritise training and education	✓	✓		✓	✓
12	Need for shared ownership across stakeholders	✓	✓	✓	✓	✓
Other Attributes						
13	Is the structure appropriate for the genre?	✓	✓	✓	✓	✓
14	Is the tone appropriate for the genre?	✓	✓	✓	✓	✓
15	Does it include citations and /or References?	✓	✓			
16	What is the word count?	593	501	474	389	261

Figure 2: Thematic overview of stage 4 output

Discussion

The results outlined above show that all essays generated from distinct GPT Chat accounts have a common attribute of good quality. This implies that ChatGPT is capable of generating high-quality and original content, hardly distinguishable from what a human will generate. For example, essay 1 was generated by User Account 1 and shows an exceptionally high level of originality, with 4% Turnitin similarity index. Conversely essays 2-6, which were also generated from User Account 1, show very low levels of originality because of very high similarity with essay 1. In other words, ChatGPT is user account sensitive and is therefore unable to generate multiple original content in response to the same (or a similar) prompt from the same account. On the other hand, essays 7 to 10 were generated from four separate user accounts, and produced significant levels of originality, respectively with 18, 19, 24 and 17% similarity index. An inspection of the similarity analysis indicates that a considerable fraction of these similarity indices was associated with the common text of the question, shared by all the essays. It is noteworthy that the tool appears able to create original content on the exact same prompt from different user accounts, or different devices. This evidence implies that ChatGPT in its current form cannot be deployed as a detector in the same manner as Turnitin.

We find common themes across all the output types, which suggests a tendency for ChatGPT's output on the same topic to be internally consistent. Except for a few variations, all outputs highlight a similar set of challenges (such as infrastructure and personnel limitations) and solutions (including increased investments and shared ownership). However, ChatGPT seems to struggle with referencing, having apparently provided only 'placeholder' references.[4] Findings from the thematic analysis also point at the ability of ChatGPT to be focused on the topic. Without exception, all outputs analysed, including the essay, report, blog post, editorial and poem, indeed talk about the status and constraints to digital transformation of the health sector in the global South. Every output also offers concrete suggestions on how to address the constraints.

Remarkably, ChatGPT wrote with a tone and structure that matches expectations about each output type. For instance, the poem was written in verses, the report had clearly defined sections while the editorial and blog posts mainly used simple language that is accessible to a general audience. The ability to stay on point while respecting genre combine to make ChatGPT – and indeed any similar AI-enabled tool – potential game changers in higher education. This has both a positive side, which studies like (Sharples, 2022) and Bibauw et al. (2022), among others, have previously discussed. A gaping gap in the literature on AI in education is what kind of changes will come with the use of AI tools to mediate assessment. In the next section we explore this by developing a conceptual framework for AI-enabled assessment.

Towards a conceptual framework for AI-mediated assessment for lifelong learning

Following on from the above discussion, we set out a conceptual framework that incorporates the capabilities of artificial intelligence into teaching and learning in higher education, while mitigating the side effects, for better student outcomes. Before elucidating this framework, we

[4] A placeholder reference is used in this sense to refer to a bogus reference that is included in a text to give an appearance of credibility. All the citations and references provided by ChatGPT in the essay and report were not found on Google Scholar or on the websites of the cited journals. The authors were found but not the works cited.

first set out two key premises in relation to the applications and implications of artificial intelligence. The first is the principle of lifelong learning in higher education within the context of preparing students for the new knowledge economy. The second is the integrated view of assessment as a process that is not limited to "baseline" knowledge testing and memorialisation, but also incorporate competence (know how) assessment and performance (show how) evaluation.

The concept of lifelong learning is not new (see, for example, Cropley and Knapper, 1983; Cryer, 1998). The modern concept of lifelong learning was introduced by Lindeman in 1926 when he criticised the additive model of formal education and instead proposed that education is a lifelong process of learning (Lindeman, 1926). The concept was subsequently introduced by UNESCO in 1949 but lost steam in the 70s and 80s before returning to the global agenda in the 1990s, in the wake of global recession, skyrocketing unemployment figures and the end of the Cold War (Volles, 2016). More recently, the emergence of the new knowledge economy and ongoing rapid changes precipitated by digital transformation, has heightened interest and sharpened the focus on the imperative of innovative pedagogy that prepares learners not only for the current state of the labour market but also capacitates them to adapt to changes and respond to opportunities in a rapidly evolving global economy. With the rapid pace of technological change, human workers are having to up-skill and re-skill themselves in order to remain relevant in existing roles, or otherwise access new opportunities. In line with the principles of lifelong learning, higher education providers are under increasing pressure to innovate teaching methods and restructure contents in response to the demands of the new knowledge economy.

In order to effectively capacitate students for lifelong learning, there is a need for a comprehensive, integrated framework of assessment that is not limited to testing the ability of students to memorise and recall taught contents, but also their capacity to apply and adapt them to dynamic, real-life situations. Again, the idea of an integrated assessment model is not new. As mentioned in the previous section, an integrated framework of clinical assessment was proposed by Miller in 1990 (Miller, 1990). In Miller's framework, the pyramid of learning outcomes and assessment begin with knowledge (know what) testing at the base, to competence (know how) assessment at the second level, performance (show how) evaluation at the third level, and action demonstration at the topmost level. An integrated framework of assessment is well aligned with the principles of lifelong learning and the imperative of a dynamic knowledge economy. Learners who know how to apply acquired knowledge to specific real-life situations are more likely able to apply their skills and competencies to similar situations or otherwise adapt or upgrade their skills to different real-life situations. Learners who have gone further to show their knowledge and skills in real life situations, say in internship, gap year or consultancy projects, would most likely have experienced and adapted themselves to a variety of practical real-life scenarios different from textbook templates. As such, they would be better prepared for different real-life situations they are likely to encounter in a post-study world of work.

While the merits of Miller's four-level framework of assessment are evident, and has been widely applied for example in clinical training, it has not achieved similar levels of adoption in other disciplines. This is on account of operational constraints associated with human resource limitations, logistical challenges and other practical difficulties inherent in, for example, implementing these in pure and non-vocational disciplines. We argue that recent, and

ongoing, advances in artificial intelligence, offer untapped potentials and opportunities to mitigate, if not altogether eliminate, these challenges. In sum, we argue that artificial intelligence can be harnessed as complimentary tools for both formative and summative assessment across three levels of assessment: knowledge (know what) testing; competence (know how) assessment, and performance (show how) evaluation. The co-option of AI tools will invariably reduce the pressure on staff time, enabling them to focus attention on other, including affective, aspects of pedagogical interventions to which human actors are best suited. We focus on the first three of Miller's levels of assessment on the basis that they are the ones most likely to be shared across a whole spectrum of disciplines. We explicate the framework as follows.

Beginning with the knowledge (know what) level of assessment, we propose that:

AI tools, such as Chatbots, can be used to support formative instant text feedback for learners (proposition 1)

AI tools, such as automated essay scoring systems, can be deployed to assess summative assessments, thereby freeing up staff time (proposition 2)

The first proposition focuses on capabilities embedded in transformer AI systems such as Chat GPT enable both formative and summative assessment of learners' knowledge. For instance, developing tools that embed AI into existing feedback systems will make them more dynamic and capable of providing a more realistic assessment of the progress of individual learners. During learning sessions, formative assessments of the future may be transformed with AI-enabled tools that deploy computer-aided quizzes that is capable of dynamically estimating individual learners' abilities and administering items that match the learner's ability (Choi and McClenen, 2020; Yang, Flanagan and Ogata, 2022a). AI-mediated summative assessment is potentially more efficient and less costly because it requires far less time commitment from teaching staff. It is therefore appealing to deploy AI tools in automatically scoring and providing feedback on assessment tasks such as essays and computer codes. Such automated assessments are established in the literature to be largely indistinguishable from human grading and offer a useful complement to the human teacher (Vittorini, Menini and Tonelli, 2021).

Moving to the second, competence (know how) level of assessment, we propose as follows:

AI-assisted, computerised adaptive feedback (CAF) can be used to provide formative, timely, personalised assistance, thereby improving learners' engagement and study habits (proposition 3)

AI-assisted computerised adaptive testing can be deployed for summative assessment that are adaptable to learners' competence and personalised learning (proposition 4)

The above set of propositions highlight the capabilities of AI tools to be deployed in assessments of competence levels. This is beyond the baseline of knowledge testing, often characterised by memorisation, recall and, at best, generation of new knowledge through the aggregation and synthesis of extant knowledge. As Chat GPT has demonstrated, AI transformers are able to synthesise extant knowledge in order to generate new knowledge, in the process rendering human learners more passive than normal. With competence outcomes, AI tools are typically co-opted in more collaborative ways by active learners, in both formative and summative processes. Computerised adaptive testing (CAT) computerised adaptive

feedback (CAF) and are prime examples of this collaborative process. Summative CATs are item-level tests that are adaptable to examinees' demonstrated ability levels, thereby providing tailored and personalised learning and assessment(Oppl *et al.*, 2017; Gardner, O'Leary and Yuan, 2021). They have been used in clinical and professional competence testing and offer promising applications in other disciplines. More recently, with the advent of versatile AI tools, computerised adaptive feedback can also be applied for formative learning and competence testing. Formative adaptive systems progressively generate items that are suited to learners' competence levels, adjust these quizzes as learners progress in relation to previously unattempted problems (Yang, Flanagan and Ogata, 2022b). It also identifies, and generates content and feedback on, items that need to be reviewed.

Finally, at the third, performance level of assessment, we propose that:

Computer serious games offer learners unlimited formative feedback opportunities in simulated real-life contexts (proposition 5)

Computer serious games offer summative assessment of situated and experiential learning through active experimentation and immersion in the game (proposition 6)

HE providers have long recognised the value of real-life situations as an important component of students' learning experience. As such, options for internship, work experience, gap year, apprenticeships, and consultancy projects have become increasingly popular across undergraduate and postgraduate programmes. They provide opportunities for learners to apply their skills and competences in real-life contexts. These offers are however resource intensive, and placements are sometimes competitive and not equally available across university programmes. The quality of the experience may also vary according to the sector or specific activities students are able to engage in, and the kind of support they are able to access.

In response to the challenges and constraints of traditional work-based learning, artificial intelligence offers unique opportunities to simulate a wide range of real-life scenarios via computer serious games. These AI tools effectively, if not perfectly, mirror dynamic real-life work situations for which static competences are not adequate. In other words, it is not sufficient to know how to deal with a specific scenario, but also to show how to engage when that specific scenario changes, as it so often does in the 21st century world of work. In effect, through interaction with the AI interfaces, learners begin to enact the process of upgrading and adapting their competences to dynamic, simulated real life scenarios, while still in formal education. This approach effectively capacitates and habituates students for lifelong learning. Formative serious games offer unlimited feedback opportunities in an iterative process of continuous learning (Ormeño *et al.*, 2019; Hainey *et al.*, 2022). Similarly, summative serious games evaluate situated and experiential learning through active experimentation and immersion in the game (Girard, Ecalle and Magnan, 2013). In combination, they provide learners with critical opportunities to learn and relearn, and to apply and adapt their skills and competences in relation to moving targets that characterises the 21st century world of work.

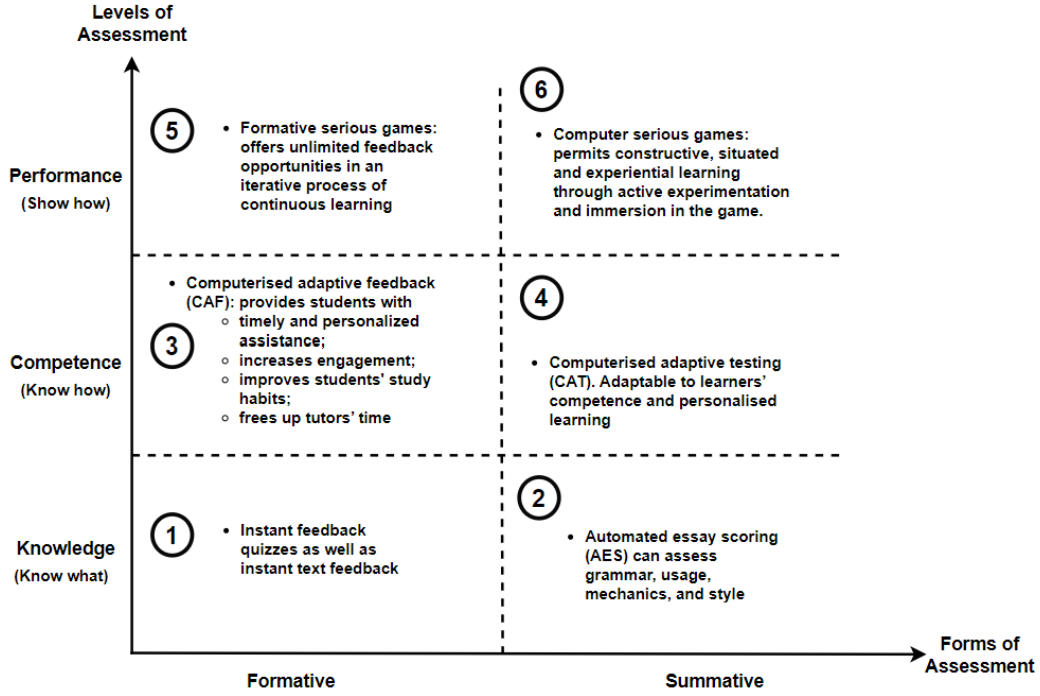

Figure 3: Integrated assessment matrix for lifelong learning: a conceptual framework

Conclusion

The application of artificial intelligence in education has received much attention, precipitated by the advent of GPT-3. In this paper, we set out to explore the implications of GPT-3 for learning and assessment in higher education. We implemented an experiment and then developed a framework based on the experimental results, in order to address two intertwined research questions. The research questions relate to the impact of GPT-3 on the evaluation of students' learning and the opportunities offered by GPT-3 to enrich learning experience in higher education.

The experiment performed on ChatGPT revealed that it can generate high-quality, original content that is hard to distinguish from human-generated content. The Turnitin similarity index of essays generated by different user accounts varies, with the first essays generated by each account having a low index, while subsequent essays have a high index. This suggests that ChatGPT is user account sensitive and cannot generate multiple original content in response to the same prompt from the same account. However, it is capable of creating original content for the same prompt from different user accounts or devices. Thematic analysis revealed common themes across different output types, indicating Chat GPT's ability to be focused on a topic and write in a tone and structure that matches expectations for the genre. However, it struggles with referencing. Based on these findings, a conceptual framework for AI-enabled assessment is proposed that incorporates AI into teaching and learning in higher education while mitigating side effects for better student outcomes. The framework is based on the principles of lifelong learning and integrated assessment. It identifies six specific domains within which AI could be applied and provides examples of such applications.

Conclusion

The capabilities of transformer AI interfaces, such as Chat GPT has sharpened the focus of HE stakeholders on the limited and limiting value of learning and assessment model that is disproportionately oriented towards knowledge testing. Knowledge creation will continue to be an important learning outcome and assessment in HE. However, in the 21st century HE, and in line with the changes and needs in the new knowledge economy, learning has to be more than the ability to create new knowledge, but also incorporate the competence to apply knowledge, and the ongoing performance of knowledge and competence driven action in real-life situations. In these regards, artificial intelligence, including transformer AI interfaces, offer endless opportunities to be co-opted into innovative curricula and assessment. AI tools can simulate real-life scenarios in which learners' competence is actioned in a dynamic iterative process that, in the same breadth, offers practically unlimited opportunities for feedback and continuous learning. In effect, rather than taking an approach of outright resistance to AI tools, higher education providers should embrace the new frontiers of opportunities presented by artificial intelligence to enrich learners' experience and enhance student outcomes. Paradoxically, this open approach will invariably empower agile HE providers to effectively curtail any challenges and dark sides of artificial intelligence.

The capabilities of transformer AI interfaces, such as Chat GPT has sharpened the focus of HE stakeholders on the limited and limiting value of learning and assessment model that is disproportionately oriented towards knowledge testing. Knowledge creation will continue to be an important learning outcome and assessment in HE. However, in the 21st century HE, and in line with the changes and needs in the new knowledge economy, learning has to be more than the ability to create new knowledge, but also incorporate the competence to apply knowledge, and the ongoing performance of knowledge and competence driven action in real-life situations. In these regards, artificial intelligence, including transformer AI interfaces, offer endless opportunities to be co-opted into innovative curricula and assessment. AI tools can simulate real-life scenarios in which learners' competence is actioned in a dynamic iterative process that, in the same breadth, offers practically unlimited opportunities for feedback and continuous learning. In effect, rather than taking an approach of outright resistance to AI tools, higher education providers should embrace the new frontiers of opportunities presented by artificial intelligence to enrich learners' experience and enhance student outcomes. Paradoxically, this open approach will invariably empower agile HE providers to effectively curtail any challenges and dark sides of artificial intelligence.

Admittedly, this study has some limitations which provide avenues for future research. First, it is possible that the performance of ChatGPT has been influenced by the choice of topic and geographical context used in the prompts. Larger studies that apply prompts on a wide range of subjects and contexts may help to shed light on this aspect. A similar case can be made for studies from different disciplinary areas. Finally, a comparative analysis where students are assigned the same written exercises as ChatGPT could provide useful insight on how future AI-mediated assessments may be designed.

References

Bibauw, S. *et al.* (2022) 'Dialogue systems for language learning: a meta-analysis', *Language Learning & Technology*, 26(1).

Checco, A. *et al.* (2021) 'AI-assisted peer review', *Humanities and Social Sciences Communications*, 8(1). Available at: https://doi.org/10.1057/s41599-020-00703-8.

Choi, Y. and McClenen, C. (2020) 'Development of Adaptive Formative Assessment System Using Computerized Adaptive Testing and Dynamic Bayesian Networks', *Applied Sciences*, 10(22), p. 8196. Available at: https://doi.org/10.3390/app10228196.

Covic, T. and Jones, M.K. (2008) 'Is the essay resubmission option a formative or a summative assessment and does it matter as long as the grades improve?', *Assessment and Evaluation in Higher Education*, 33(1), pp. 75–85. Available at: https://doi.org/10.1080/02602930601122928.

Crook, C. and Nixon, E. (2021) 'How internet essay mill websites portray the student experience of higher education', *Internet and Higher Education*, 48. Available at: https://doi.org/10.1016/j.iheduc.2020.100775.

Cropley, A.J. and Knapper, C.K. (1983) 'Higher Education and the Promotion of Lifelong Learning', *Studies in Higher Education*, 8(1), pp. 15–21. Available at: https://doi.org/10.1080/03075078312331379081.

Cryer, P. (1998) 'Transferable Skills, Marketability and Lifelong Learning: The particular case of postgraduate research students', *Studies in Higher Education*, 23(2), pp. 207–216. Available at: https://doi.org/10.1080/03075079812331380394.

Dale, R. (2021) 'GPT-3: What's it good for?', *Natural Language Engineering*. Cambridge University Press, pp. 113–118. Available at: https://doi.org/10.1017/S1351324920000601.

Dixson, D.D. and Worrell, F.C. (2016) 'Formative and Summative Assessment in the Classroom', *Theory into Practice*, 55(2), pp. 153–159. Available at: https://doi.org/10.1080/00405841.2016.1148989.

Dunn, K.E. and Mulvenon, S.W. (2009) 'A Critical Review of Research on Formative Assessments: The Limited Scientific Evidence of the Impact of Formative Assessments in Education', *Practical Assessment, Research, and Evaluation*, 14, p. 7. Available at: https://doi.org/10.7275/jg4h-rb87.

Gardner, J., O'Leary, M. and Yuan, L. (2021) 'Artificial intelligence in educational assessment: "Breakthrough? Or buncombe and ballyhoo?"', *Journal of Computer Assisted Learning*. John Wiley and Sons Inc, pp. 1207–1216. Available at: https://doi.org/10.1111/jcal.12577.

Girard, C., Ecalle, J. and Magnan, A. (2013) 'Serious games as new educational tools: How effective are they? A meta-analysis of recent studies', *Journal of Computer Assisted Learning*, 29(3), pp. 207–219. Available at: https://doi.org/10.1111/j.1365-2729.2012.00489.x.

Goss, H. (2022) ‘Student Learning Outcomes Assessment in Higher Education and in Academic Libraries: A Review of the Literature’, *Journal of Academic Librarianship*, 48(2). Available at: https://doi.org/10.1016/j.acalib.2021.102485.

Guo, K., Wang, J. and Chu, S.K.W. (2022) ‘Using chatbots to scaffold EFL students’ argumentative writing’, *Assessing Writing*, 54. Available at: https://doi.org/10.1016/j.asw.2022.100666.

Hainey, T. *et al.* (2022) ‘Serious Games as Innovative Formative Assessment Tools for Programming in Higher Education’, in *Proceedings of the 16th European Conference on Games Based Learning*, pp. 1–10.

Lindeman, E. (1926) *The meaning of adult education*. 1st edn. New York: New Republic.

McNamara, D.S., Crossley, S.A. and Roscoe, R. (2013) ‘Natural language processing in an intelligent writing strategy tutoring system’, *Behavior Research Methods*, 45(2), pp. 499–515. Available at: https://doi.org/10.3758/s13428-012-0258-1.

Miller, G.E. (1990) ‘The assessment of clinical skills/competence/performance’, *Academic Medicine*, 65(9), pp. S63-67. Available at: https://journals.lww.com/academicmedicine/Abstract/1990/09000/The_assessment_of_clinical.45.asp (Accessed: 28 January 2023).

Oppl, S. *et al.* (2017) ‘A flexible online platform for computerized adaptive testing’, *International Journal of Educational Technology in Higher Education*, 14(1). Available at: https://doi.org/10.1186/s41239-017-0039-0.

Ormeño, E. *et al.* (2019) ‘Towards a formative instrument to evaluate user experience in virtual reality serious games’, in *ACM International Conference Proceeding Series*. Association for Computing Machinery, pp. 53–57. Available at: https://doi.org/10.1145/3364138.3364152.

Rawlusyk, P.E. (2018) ‘Assessment in Higher Education and Student Learning’, *Journal of Instructional Pedagogies*, 21, p. 1. Available at: http://www.aabri.com/copyright.html.

Sharples, M. (2022) ‘Automated Essay Writing: An AIED Opinion’, *International Journal of Artificial Intelligence in Education*. Springer, pp. 1119–1126. Available at: https://doi.org/10.1007/s40593-022-00300-7.

Vittorini, P., Menini, S. and Tonelli, S. (2021) ‘An AI-Based System for Formative and Summative Assessment in Data Science Courses’, *International Journal of Artificial Intelligence in Education*, 31(2), pp. 159–185. Available at: https://doi.org/10.1007/s40593-020-00230-2.

Volles, N. (2016) ‘Lifelong learning in the EU: changing conceptualisations, actors, and policies’, *Studies in Higher Education*, 41(2), pp. 343–363. Available at: https://doi.org/10.1080/03075079.2014.927852.

Wang, X. *et al.* (2023) ‘What matters in AI-supported learning: A study of human-AI interactions in language learning using cluster analysis and epistemic network analysis’, *Computers and Education*, 194. Available at: https://doi.org/10.1016/j.compedu.2022.104703.

Wang, Z. (2022) ‘Computer-assisted EFL writing and evaluations based on artificial intelligence: a case from a college reading and writing course’, *Library Hi Tech*, 40(1), pp. 80–97. Available at: https://doi.org/10.1108/LHT-05-2020-0113.

Yang, A.C.M., Flanagan, B. and Ogata, H. (2022a) ‘Adaptive formative assessment system based on computerized adaptive testing and the learning memory cycle for personalized learning’, *Computers and Education: Artificial Intelligence*, 3, p. 100104. Available at: https://doi.org/10.1016/j.caeai.2022.100104.

Yang, A.C.M., Flanagan, B. and Ogata, H. (2022b) ‘Adaptive formative assessment system based on computerized adaptive testing and the learning memory cycle for personalized learning’, *Computers and Education: Artificial Intelligence*, 3, p. 100104. Available at: https://doi.org/10.1016/j.caeai.2022.100104.