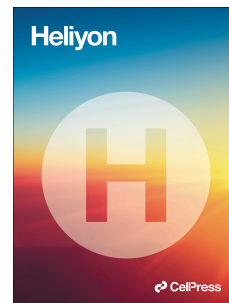


Journal Pre-proof

Is AI changing learning and assessment as we know it? Evidence from a ChatGPT experiment and a conceptual framework

Oluwaseun Kolade, Adebowale Owoseni, Abiodun Egbetokun



PII: S2405-8440(24)01984-4

DOI: <https://doi.org/10.1016/j.heliyon.2024.e25953>

Reference: HLY 25953

To appear in: *HELIYON*

Received Date: 29 August 2023

Revised Date: 27 January 2024

Accepted Date: 5 February 2024

Please cite this article as: , Is AI changing learning and assessment as we know it? Evidence from a ChatGPT experiment and a conceptual framework, *HELIYON* (2024), doi: <https://doi.org/10.1016/j.heliyon.2024.e25953>.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2024 Published by Elsevier Ltd.

Is AI changing learning and assessment as we know it? Evidence from a ChatGPT experiment and a conceptual framework

Oluwaseun Kolade^{a1} (corresponding author), Adebowale Owoseni^b, and Abiodun Egbetokun^c

^a Sheffield Business School, Sheffield Hallam University, Sheffield UK

^b School of Computer Science and Informatics, De Montfort University, Leicester UK

^c Leicester Castle Business School, De Montfort University, Leicester UK

Abstract

ChatGPT, a state-of-the-art chatbot built upon Open AI's generative pre-trained transformer, has generated a major public interest and caused quite a stir in the higher education sector, where reactions have ranged from excitement to consternation. This paper therefore examines the potential impact of ChatGPT on learning and assessment, using the example of academic essays, being a major form of assessment with widespread applications of ChatGPT. This provides an opportunity to unpack broader insights on the challenge of generative AI's to the relevance, quality and credibility of higher education learning in a rapidly changing 21st century knowledge economy. We conducted a quasi-experiment in which we deployed ChatGPT to generate academic essays in response to a typical assessment brief, and then subjected the essays to plagiarism checks and independent grading. The results indicate that ChatGPT is able to generate highly original, and high quality, contents from distinct individual accounts in response to the same assessment brief. However, it is unable to generate multiple original contents from the same account, and it struggled with referencing. The discussion highlights the need for higher education providers to rethink their approach to assessment, in response to disruption precipitated by artificial intelligence. Thus, following the discussion of empirical data, we propose a new conceptual framework for AI-assisted assessment for lifelong learning, in which the parameters of assessment extend beyond knowledge (know what) testing, to competence (know how) assessment and performance (show how) evaluation.

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 [1]. Chatbots-software applications designed to simulate human-like conversations by responding to voice or text inputs- are demonstrably effective as pedagogical tools. They offer 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 [2].

In recent years, rapid advances in artificial intelligence have led to the emergence of generative pre-trained transformer a state-of-the art autoregressive language models which offerings and capabilities far supersedes previous models of chatbots. With 175 billion parameters at its

¹ Corresponding author: Sheffield Business School, Sheffield Hallam University, Sheffield UK, Email: seunkolade2014@gmail.com (Oluwaseun Kolade)

command, ChatGPT is one of the largest and most powerful language processing AI models available [3]. With its vast and versatile capabilities, ChatGPT has been used to produce academic essays, technical reports, comedy scripts and poetry, to mention a few. The power of ChatGPT lies in its unprecedented capabilities to mimic human produced texts [4]. Compared to other existing LLMs, ChatGPT has exceptional conversational task-processing abilities [5]. 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 was specifically trained with over 570 gigabytes of data [6], moreover, the self-supervised learning approach used in the development of ChatGPT 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 [7]. 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, 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 [8].

As the enormous capabilities of CHATGPT 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 [9, 10]. 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 [11]. 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.

ChatGPT is a disruptive game changer that further complicates and exacerbates the intractable challenge of essay milling, but also potentially offer new and promising pathways to learning and assessment. First, the emergence of ChatGPT 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 generative AI, such as ChatGPT, on the evaluation of students’ learning? Secondly, what new opportunities are offered by generative AI to enrich students’ learning experience? The first question is empirical, the second conceptual. The empirical component focuses on evaluation 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 generative AI 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 ChatGPT, and analytical procedures using Turnitin and thematic analysis, 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

Learning and assessment in higher education

Historically, assessments have been used in college and universities evaluate and certify students' learning [12]. 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 [13]. 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 [14]. As such, they are typically teaching centred and used to establish learners' academic progress based on some established criteria [15].

Both summative and formative assessments have been subject to criticisms since they were introduced and have come under more intense scrutiny in recent years, in the wake of the new knowledge economy underpinned by information explosion and digital transformation. Formative assessments have been criticised for often failing to deliver on their lofty promises of active student participation and continuous supportive feedbacks through a wide range of tasks [12, 15]. Conversely, many scholars have argued that summative tests promote memorisation rather than understanding and application that are associated with deep learning [12, 13]. This approach, it is argued, is not adequate for preparing students for a rapidly changing knowledge economy in which they are expected to function.

In furtherance of, and response to, critiques of formative and summative assessments, several scholars have proposed a third purpose of assessment. The third purpose, it is suggested, is to foster learning through life- beyond the traditional remits of summative and formative assessments [16-18]. Learning through life is socially situated and dynamically embedded in everyday tasks and real-life contexts, in contrast with the abstract character of classroom learning in higher education [16]. To prepare students for lifelong learning, scholars have argued that assessments should be attuned to the dynamic and rapidly changing contexts of the 21st century workplace [19]. It should also incorporate active student involvement, including through peer assessment and self-assessment; and it should embrace the idea of feedback as "feedforward", where feedback is not merely a review and grading of completed assessments, but incorporates insights and guidance that learners can take forward into future tasks and situations [18, 19].

This explication of the type and purpose of assessment is important within the context of practitioners and public interest on the disruptive impact of generative AI in the HE sector and amid growing calls to rethink the future and models of assessment. The followings sections outline perspectives from the extant literature on the potential roles of AI in formative and

summative assessment, in general, followed by a critique of their implications for essays, in particular.

Use of AI in learning and assessment

Digital tools, including audio-visual, computer and telematic tools that enabled remote and asynchronous learning, have been applied for a long time in higher education [20]. However, remote and asynchronous teaching and learning have become more popular since the COVID-19 pandemic [21]. These modes of delivery underlie the digital transformation that is now taking place in education [22]. 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 ChatGPT can generate curriculum content, fix bugs in computer codes and write complex passages that compare to human output [23].

The pedagogical potential of AI tools like ChatGPT is increasingly recognised among researchers and practitioners. A meta-analysis by [24] indicates that AI chatbots are strikingly effective for language learning. In a systematic review of 74 publications on chatbots' application in education, [25] 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 [26]. Moreover, algorithmic bias may limit the objectivity of AI tools, a limitation that users are unlikely to be aware of [27].

AI tools also have significant potentials for applications in assessment. Prior to ChatGPT, purpose-built AI tools have been successfully applied for assessment in the context of research and education [28]. For instance, [27] 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, [29] reported that an AI-enabled system performed better than human teachers in terms of feedback effectiveness and impact on students' learning ability. Similarly, [30] 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" [31, pg. 44]. For this reason, today's assessment tests in higher education still follow a tradition that started in the 19th century [32] which relies on written essays and reports as the gold standard. The next section delves into the dominance of essays as a means of assessment in higher education and how it could potentially be impacted by AI, especially LLMs like ChatGPT. of essays as a means of assessment in higher education and how it could potentially be impacted by AI, especially LLMs like ChatGPT.

Essays as a method of assessment and the implications of AI

The written essay is one of the most prominent formats of summative assessments in higher education. It is conceived to evaluate knowledge comprehension and understanding, and to enable new knowledge construction emanating from learners' critical thinking skills. 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 [33]. Essays are also known to reduce students' pressure for rote learning in contrast to traditional examinations which often promote rote learning [34] (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 [35]. Perhaps for this reason, AI-enabled tools such as OpenEssayist [36] and Writing Pal [30] have been developed specifically to support students in writing essays.

However, with the rapid evolution of artificial intelligence and the emergence of ChatGPT, stakeholders are now questioning the continued use of essays in its current format to evaluate students' learning outcomes. Despite being the most widely used form of summative assessment in higher education today, written essays may be viewed as inherently weak in many ways. First, essays may undermine effective learning because of the risk of collusion and copying. Moreover, essays are alleged to have limited objectivity and high susceptibility to unethical practices such as plagiarism and wholesale ghost writing [23, 37]. In response to these challenges, plagiarism checks have become a core feature of essay assessment in higher education. But 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. For instance, [4] showed that ChatGPT-3.5 and ChatGPT are capable of consistently generating original and authentic content. 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.

In the face of rising AI possibilities, it remains to be seen how teaching and assessment models will evolve. CHATGPT 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 [23]. A recent study by [38] found that existing tools such as OpenAI, Writer, Copyleaks, GPTZero, and CrossPlag which are designed to detect AI-generated content tend to fail when confronted with the most advanced GPT version. In another assessment involving five of the latest AI content detection tools (GPTZero, OpenAI Text Classifier, Writer.com's AI Content Detector, Copyleaks AI Content Detector, and Giant Language model Test Room), [39] concluded that these tools "seem not yet fully ready to accurately and convincingly detect AI-generated content from machine-generated texts in different contexts." In a world where cheating on written essays is already high [37] this poses an even greater challenge as LLMs' text generating capabilities continue to advance.

In summary, 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. Students need to demonstrate an ability to produce new knowledge, not merely reproduce what has been given to them by teachers. In the absence of that, the certification of learning achievement becomes

problematic. Following a thorough evaluation of the content generation capabilities of ChatGPT and the risk that this poses for the effectiveness of essays as a form of assessment, we propose a new conceptual framework for AI-enabled assessment. The framework builds upon a framework of assessment originally proposed for clinical training by [40]. In the framework, 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. Thus, in this paper, we note that ChatGPT 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

The approach

The quasi-experiment approach was considered an appropriate method for investigating the impact of GPT-3.5 on student evaluation and the new opportunities it offers for enhancing students’ learning experience. As an exploratory research design, the quasi-experiment combines the aspects of a laboratory experiment with consideration of the social implications in real-world scenarios [41]. This type of design is commonly used in educational, medical, and social science research to study the effects of interventions or programs on outcomes [41-43].

The procedure and analysis

The experiment involved twenty postgraduate (Masters) students recruited through purposive sampling strategy. The main criteria for selecting the participants include awareness of ChatGPT, personal drive to try ChatGPT, and current studentship in the social sciences. Geographically, five of the students who produced essays 1 to 10 live and study in the UK, while the remaining participants are a mix of students from the UK and Nigeria. The motivation for spreading participation across two countries is to improve the validity of the experiment. The research was designed to unfold in seven stages, as shown in Figure 1 below. The figure illustrates how the experiment progressed from "start" to "end".

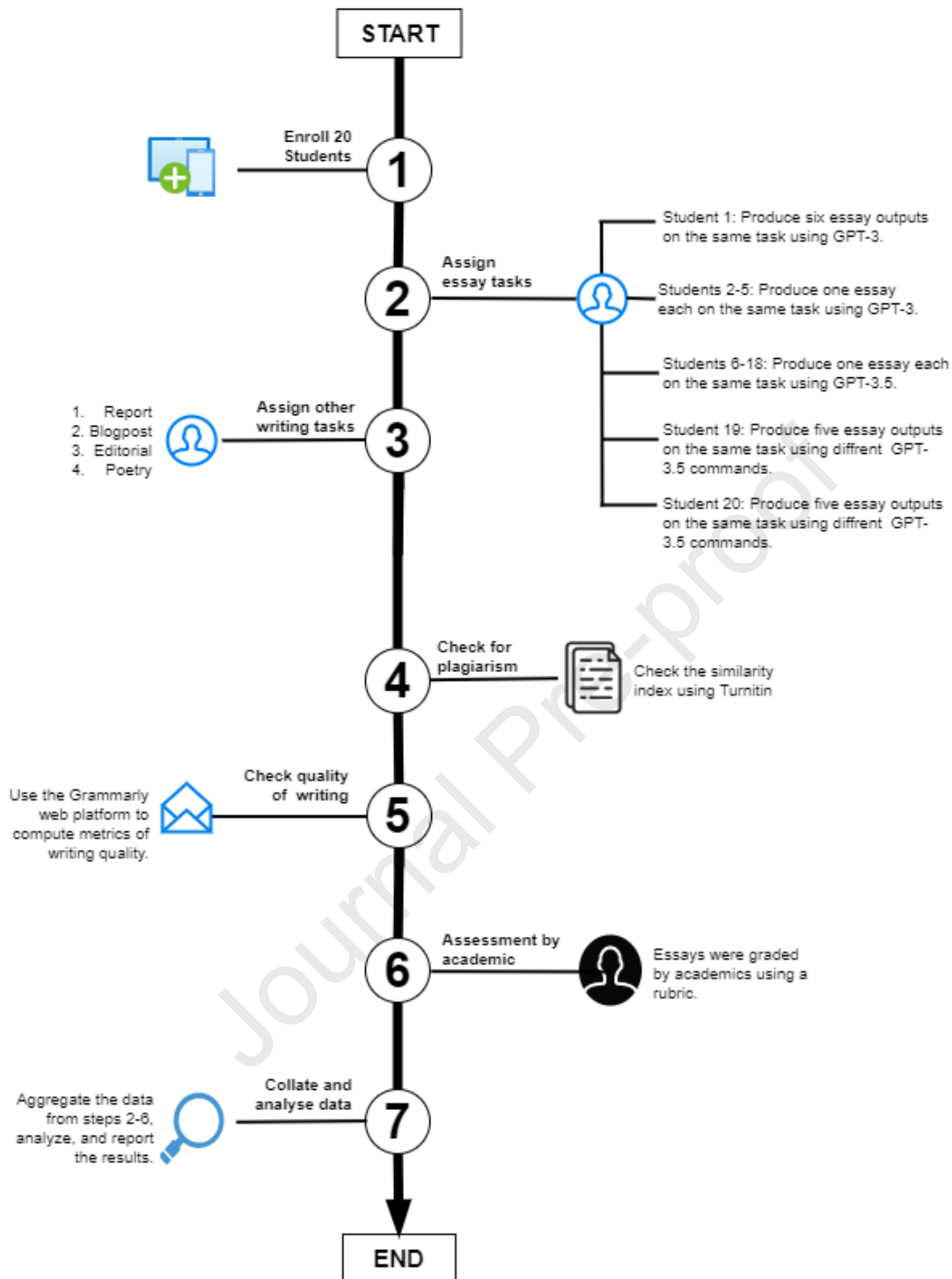


Figure 1. Quasi-experiment procedure

Stage 1: Enrolled 18 participants who met the recruitment criteria. Provided a clear, easy-to-understand description of the task, including its timeline. Explained how participants would complete and submit their tasks via email.

Stage 2: Assigned essay writing tasks required students to write an essay on the digital transformation of health sectors in the global south, with suggestions for improving digitally-enabled healthcare delivery. The tasks varied between groups. First, Student-1 used their ChatGPT 3.0 account to produce and submit six essays within five days. The second group,

Students 2-5, each submitted one essay using ChatGPT 3.0. A third group of 13 students had a similar task but used ChatGPT 3.5, generating one essay each. The final group, consisting of two students, created five essays each using different commands on ChatGPT 3.5. The difference between this group and the first student is that the latter used the same command to generate five essays with ChatGPT 3.0, while the former used different commands for their five essays on ChatGPT 3.5. In summary, this stage involved 20 students generating a total of 33 essay outputs. Stage 3: Students were instructed to use the same ChatGPT accounts to create four different types of written pieces, all centred on the topic - “*The digital transformation of the health sectors in the global south.*” or each piece, the instructions given to ChatGPT were modified to suit the specific genre required. The genres for this task included a report, a blog post, an editorial, and a poem.

Stage 4: The essays received from participants in the second stage were downloaded and tagged with unique identifiers. These identifiers anonymized the participants and distinguished each essay. This process yielded 23 essays from 18 participants. The essays were then uploaded to Turnitin to test for plagiarism. The similarity index score for each essay was noted and recorded

Stage 5: Similar to the previous stage, the essays were uploaded to the Grammarly web platform to test the quality of writing. This process generated metrics that offered insights into word count, word length, unique words, and rare words. These metrics are indicative of the readability, vocabulary, and overall quality of the essays from a system perspective. All metrics were noted and recorded.

Stage 6: This stage involved human assessments of the essays. Two academics were recruited as human assessors to evaluate the essays using a standard rubric (refer to Appendix 1). The two assessors are PhD holders and university lectures in business and digital innovation with more than 10 years aggregate teaching experience. Moreover, the assessors were within the professional network of the researchers as it is more convenient and relatable. The average score and summarized textual feedback from the assessors were noted and recorded.

Stage 7: The dataset from stages 2 to 6 was gathered and analysed to understand the impact of generative AI, such as ChatGPT, on student learning and assessment. We also considered how generative AI can enhance the learning experience. The analysis included a thematic evaluation of comments from human assessors, which was cross-referenced with the Turnitin similarity index, the Grammarly writing quality scores, and the rubric-moderated human assessment score. Additional metrics providing context to the analysis included the word count of the essays, as well as the versions and specific student accounts used to generate the essays.

We would like to clarify that the reports from two candidates were excluded from the experiment because their essays were less than 300 words and were only outlines, which cannot be considered as valid outputs. These two candidates were not included as part of the enrolled participants. Moreover, we initially attempted to use CopyLeaks AI detector but found its detection imprecise and too generic and inadequate for direct comparison with Turnitin.²

² This does not support the conclusion that AI detection tools are now able to reliably detect AI-generated text. Instead, it only reflects the fact that detection tools are getting better although they continue to lag the latest LLMs [38] (Elkahat et al, 2023).

Results and Discussion

Results

Table 1 contains the Turnitin similarity index, the Grammarly writing quality scores, and the rubric-moderated human assessment score. The Turnitin similarity index of the essays ranged from 4% to 99%. A low index was observed for the first essays generated by the first participant (St 1), which generated 6 essays, the similarity index increased significantly after the first essay, from 4% to 86% and then to 99%. Other essays written by the participants produced less than 25% similarity index except St 11 which has 31%, which suggest that Turnitin is not effective in identifying plagiarised contents written with using multiple user accounts or profiles. Although the comments from the human assessor clearly identified plagiarised content similar to Turnitin, human's assessment of the essay is largely passive resulting in high scores, which corroborate Grammarly's view of the quality of writing and readability. User IDs 19 and 20 were encouraged to use different prompts to generate five versions of the essay in response to the same assignment brief. It must also be noted that, unlike the first cohort who used ChatGPT3.0, these users employed ChatGPT3.5. Their outputs indicate that using different prompts, in combination with new capabilities of ChatGPT3.5, generated outputs with low similarity scores from the same user account. However, we also note that the word counts, especially in the 2nd to 5th iterations of the prompts, is significantly lower than the required word count of 1500 words. This probably indicates that ChatGPT was struggling a bit, from the same user account, to match the required word count for the same assignment brief, perhaps in the search for new contents and phrasings. Figure 2 provides a graphical summary of the results. Analysis with Grammarly indicate improved vocabulary and readability in essays generated using ChatGPT 3.5 compared to those created with ChatGPT 3.0. However, the independent expert assessment of these essays showed variation among participants 7 to 18, with scores ranging from a high of 77% to a low of 40%. This reflects a difference between grammatical finesse and academic substance in relation to the assessment brief.

Table 1: Analysis of user accounts and quality of essays

Candidate ID	Essay ID	Turnitin Similarity Index	Word Count	Overall Writing Quality Score (Grammarly)	Overall Assessment Score (Human)	Summary of human feedback
S1	E1	4	614	80	67	This is a satisfactory write up. The student gave a good understanding with some depth and clarity although there are rooms for improvement. The student failed to give the definition and overview of the topic, zero case study, few word count and the conclusion is a repetition of the solutions already highlighted
S1	E2	86	551	80	64.5	Satisfactory write-up with room for improvement. (This write up is exactly the same as Candidate 3's (ChatGPT Student3) submission)
S1	E3	99	546	80	64	Satisfactory write up. The student can improve on this area; Brief definition and overview, give us a case study, elaborate more on the challenges, benefit and future prospect. (The candidate has a clear understanding of the topic. He/She was able to present key areas in the health sector where digital transformation can be harnessed and also came up with challenges that needs to be addressed to fully harness the potentials in the global South. However, there are no empirical evidences stated as expected in an academic paper. The write up is however incomplete.)
S1	E4	99	543	82	64.5	A good write up, a better modification is needed on the structure , general overview, case study, and the benefit (This write up is exactly the same as Candidate 3's (ChatGPT Student3) submission)
S1	E5	88	405	83	64.5	A satisfactory essay (This write up is exactly the same as Candidate 3's (ChatGPT Student3) submission.)
S1	E6	97	408	80	64.5	Satisfactory work, an improvement is needed. (This write up is exactly the same as Candidate 3's (ChatGPT Student3) submission. My overall comment is that the essay is purely copy and paste from CHAT GPT.)
S2	E7	18	474	79	63	A very good write up with insight on advantages of the digitization of the health sectors in the global south. The conclusion is great as well. It is however under the word limit
S3	E8	19	484	82	61	Write up similar to student 7. A good write up but again falls under the word limit
S4	E9	24	428	80	57	The standard of writing is good but the essay is under the word count
S5	E10	17	397	83	55	The essay strongly addresses the assignment brief but is under the required word count
S6	E11	5	941	97	56	The understanding of the topic was observed, alongside research and literature to back statements of fact. And the conclusion is synchronised.

S7	E12	9	837	98	62	The understanding of the topic was observed, however, there were no literature to back any of the statements of fact. And the conclusion is not in tandem with the highlighted problems.
S8	E13	6	513	97	40	The candidate did not provide details on the context of Global South (countries, and their specifics). There is no clear flow among the different sections. But seems the candidate was trying to list the challenges associated with the health sector in Global South and the potentials of digital transformation on health sector in Global South.
S9	E14	0	638	97	67	The candidate has presented a logical flow of information on the subject matter and a good understanding of the topic, howbeit with little or no empirical evidence.
S10	E15	1	940	97	40	The candidate did not provide background information on the context of Global South (countries, and their specifics). Though the candidate tried to explore areas such as the emergence of Health information systems, challenges associated with the system, and how the challenges can be surmounted, there is however no clear flow among the different sections The write-up does not have logical flow and seems to be largely from a Generative AI tool.
S11	E16	31	306	97	43	The candidate only suggested areas to focus on in an effort to digitally transform the health sector of Global South countries. There is no introduction of the subject matter and the write-up lacks clear understanding and critical analysis of the subject.
S12	E17	8	962	92	68	The candidate presented a logical flow of ideas and shows good understanding of the subject matter. Although, there is hardly any evidence of citation. Seems a good portion is from ChatGPT or other Generative AIs.
S13	E18	10	971	97	65	The candidate has an impressive introductory section to the topic but failed to consolidate on this with facts and citing of relevant literatures and examples all through the essay. The write-up only defines or describes digital transformation key words.
S14	E19	15	1475	94	77.6	This candidate displayed exceptional understanding of the topic with detailed description of all terminologies associated with it. The candidate also gave the importance of each terms discussed to the eventual adoption of digital transformation of health sector in the Global South.
S15	E20	0	904	98	56.2	The candidate failed to give detailed introduction of the topic, its terms and how it relate to the health sector in the Global South. This showed that the candidate did not understand the topic or not a good writer. The candidate started with challenges and opportunities without some background of the topic.
S16	E21	11	1097	97	64.2	This candidate shows good understanding of the topic with a better writing and presentation. The candidate however failed to make a clear distinction among the sections. Hence there are lots of repetition in the write-up.

S17	E22	7	1422	86	75.8	This candidate displayed excellent understanding of the topic with detailed introduction of the topic. The candidate citing of relevant literatures and examples where required. Important terminologies were also discussed. The candidate writing style is also impressive.
S18	E23	11	416	83	58	The candidate gave the definition and an in-depth key aspect of the digital transformation in the health sectors of the Global South. Unfortunately, the holistic objectives were not achieved or looked at in this essay. The essay is also under the required word count.
S19	E24	15	920	96	62	The essay is well written, with clear points in response to the assessment task. The essay raises good and relevant points but no evidence is provided to support claims made and there is little to no critical analysis.
S19	E25	8	821	92	62	This would be an excellent essay if stayed focused on the topic, but it jumps around repeatedly from the assigned topic to a broader one. In addition to its coherence and clear structure, the content appears well thought-out and reflects some level of critical reasoning. However, there is only limited use of evidence to support the analysis. Where examples are provided, there is no reference or data to support the validity or relevance of these examples.
S19	E26	7	792	90	67	The candidate provides a well-structured discussion of relevant issues as required by the assessment task. However, there is not enough background information to help the reader understand why the specific issues were chosen. Overall, some evidence of research is present but it lacks sufficient depth and critical reflection.
S19	E27	6	854	96	72	This candidate demonstrates excellent understanding of the topic and provides a very coherent discussion. The essay could benefit from some visual aids, for example, a table that catalogues the chronology of digital transformation of the health sector in the Global South.
S19	E28	10	739	98	56	The candidate's understanding of the topic is apparently weak. In what is supposed to be a discussion of historical evolution, the essay barely mentions any dates or time periods. There are also several evidence gaps - for example the claim of a recent surge in telehealth and virtual care platforms is not supported with any data or evidence. Overall, this is a passable effort.
S20	E29	4	771	99	40	This candidate offered no discussion whatsoever but only provided a set of bullet points in each section. The sections and the points listed are, however, relevant.
S20	E30	3	814	96	66	Beginning with a strong introduction that set the tone for the rest of the essay and clearly highlight what the reader should expect, this candidate presents a decent discussion of the current status, benefits and recommended actions for digital transformation of the health system in the Global South. However, in the absence of supporting data, especially on the status quo, the analysis in the essay is deficient. There is scope to elaborate on this aspect in particular considering that the essay is well below the word limit.
S20	E31	2	953		65	This is a really good essay but it falls just short of being excellent particularly because it omits an overview of the current status of digital transformation of the health sector in the context under discussion. A major

						strong point of the essay, however, is the section where it discusses the drivers of digital transformation, although this is essentially a list.
S20	E32	2	597		55	The content of the essay suggests that the candidate understood the task and has a good idea of the right content to include. However, much of the discussion is superficial; the sections are limited in scope as if this was a rushed work. The discussion of the key drivers is quite good but there is scope for more evidence use throughout the essay to strengthen the content.
S20	E33	12	681		50	What this submission has in content, it lacks in organisation. The sequence goes from one descriptive section to another and then the candidate concludes with a paragraph that only partially connects to the rest of the content. Overall, the write-up lacks depth.

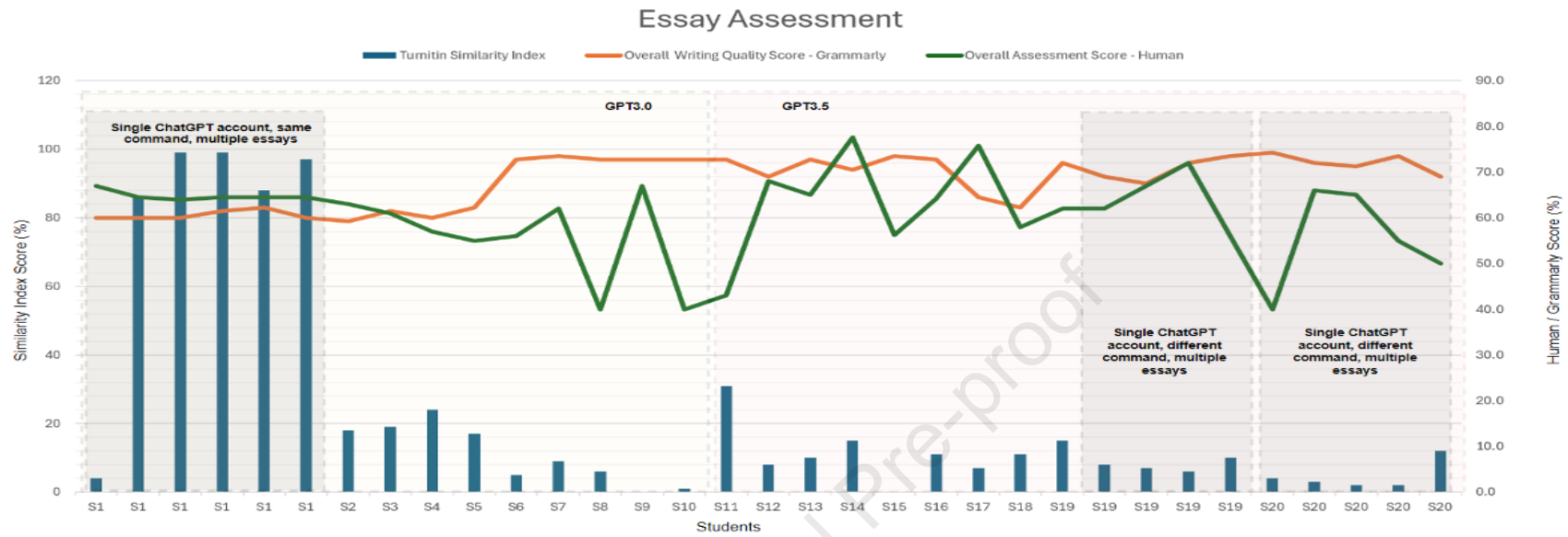


Figure 2: Overview of essay assessment

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 essay identified significant progress made in the digital transformation of the health sector in the global south. It yet identified the several challenges that need be surmounted. The specific means of surmounting the challenges via collaborative efforts between the government and private sector were mentioned. No citation of references consulted was sighted (Feedback on Essay 1)

The student wrote the essay in a language that could be understood by a wide array of readers. The essay highlighted: the merits; specific example of the digital transformation expected and barriers while mentioning ways to improve digital healthcare delivery. No citations provided. (Feedback on Essay 7)

The specific ways digitally-enabled healthcare delivery could benefit people the people of the global south with specific mention of the chronic diseases of the people there were mentioned in the essay. Specific mention was made of ways to improve the digitally-enabled healthcare delivery in the global south. No citations were provided (Feedback on Essay 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 theoretical grounding and 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...– Feedback on Essay 9

The essay addressing some of the challenges present in the global south when addressing challenged with health care through digital transformation. In addition, presents a specific technology-mobile health that can be used to deliver digital health, it would have helps to provide some examples of exactly how this can be implanted such as using USSD etc. Overall, the essay follows a logical structure with an introduction to the subject/focus and an argument for the importance of digital transformation and challenges. Also has a body and conclusion that are clearly stipulated. The essay does not draw on theory e.g. diffusion theory or even absorptive capacity theory to show how these digital health technology can be adopted in the global south... (Feedback on Essay 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 3 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 criterion, 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: Instructions given to ChatGPT and researchers' observations

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: <i>Kroneman, M., Brouwer, W., & Scott, A. (2016). Digital health in low- and middle-income countries: from promises to impact. Lancet, 387(10022), 1040-1050.</i>
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>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.</i> "
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 editorial is for a local newspaper .	In comparison to other genres, the message was more focused on the problems.
5	<i>Please write poetry for me on "the digital transformation of the health sectors in the global south".</i>	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.

	<i>The poetry will be presented at the WHO conference</i>	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 3: 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.

The independent feedback on the essays underline the limitation of Chat GPT with theory and references, but also its struggles with specific case examples and contexts to illustrate its arguments. In the case of the assessment brief used in the current study, its generic wording in relation to the global South masks this limitation somewhat, compared with say a brief focusing on a more specific context or case study.

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.³ 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.

It is noteworthy that 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 and negative side, which studies like [23] (Sharples, 2022) and [24] Bibauw et al. (2022), among others, have previously discussed. For instance, using fairly complex prompts in several iterations, an average user will manage to generate a high-quality essay that could easily pass for human work. This is more likely as the capabilities of LLMs like ChatGPT advance and detection tools continue to trail them. While it may be argued that an expert could spot AI-generated output upon close scrutiny, the sheer amount of time this will take and the associated additional workload implies that a better approach is required to enrich students' learning in a world with LLMs. In this regard, 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 empirical 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 (see Figure 4). The framework addresses the second research question set out in the introduction of this paper, that is, what new opportunities are offered by generative AI to enrich students' learning experience? This is an important question in light of the implications of LLMs highlighted by the above

³ 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.

empirical results. Before elucidating this framework, we 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, [44, 45]). 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 [46]. 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 [47]. 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 [40]. 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 the framework 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 [48, 49]. 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 [50].

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 [51, 52]. 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 [53]. 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 [54, 55]. Similarly, summative serious games evaluate situated and experiential learning through active experimentation and immersion in the game [56] 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.

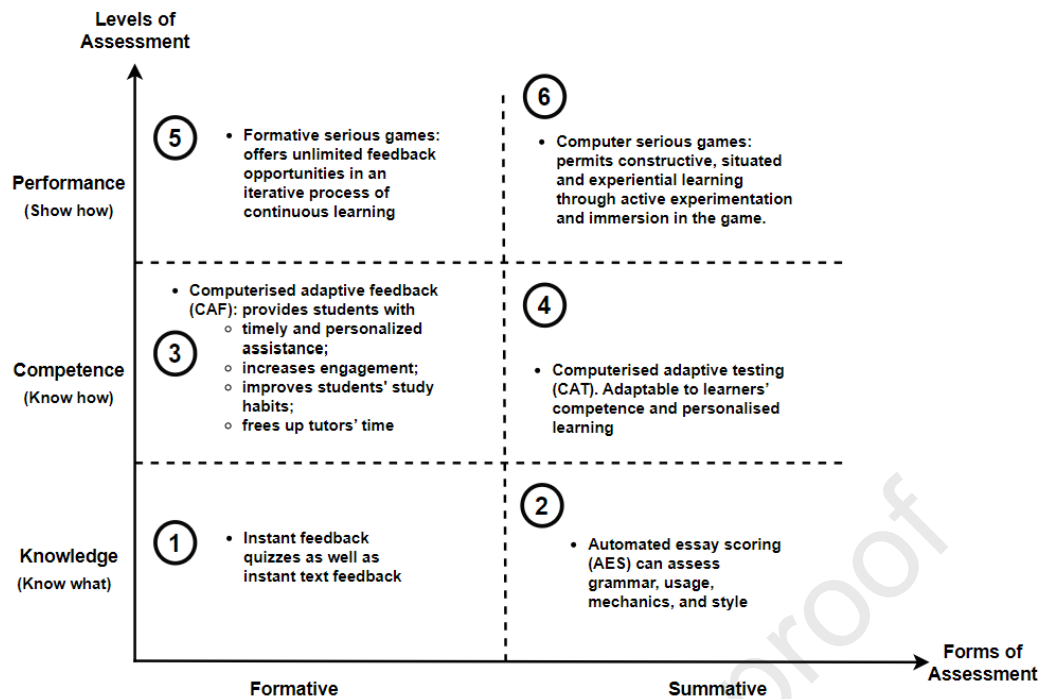


Figure 4: Integrated AI-assisted assessment matrix for lifelong learning: a conceptual framework

Conclusion

The application of artificial intelligence in education has received much attention, precipitated by the advent of ChatGPT. In this paper, we set out to explore the implications of CHATGPT 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 ChatGPT on the evaluation of students' learning and the opportunities offered by ChatGPT 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 ChatGPT'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. Operationalising this framework will require systemic changes within HE and this requires an open mind as well as a willingness to experiment on the part of HE

policymakers and decision makers. Stakeholder resistance may pose a significant challenge to operationalizing the proposed framework, as some of the required changes may be unfamiliar and uncomfortable, leading to inertia. However, we leave it to future empirical studies to trial the framework, identify more specific challenges and the relevant mitigation strategies.

The capabilities of transformer AI interfaces, such as ChatGPT has sharpened the focus of HE stakeholders on the limited and limiting value of a 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. It must also incorporate the competence to apply knowledge, and create opportunities for 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. In particular, future studies should explore the capabilities of ChatGPT in response to case study questions and discussion of primary data. A similar case can be made for studies from different disciplinary areas. In addition, 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. This study used fairly straightforward prompts and results were generated from a single run of the prompts, without any iterations. Studies that use more complex prompts and allow for several rounds of conversational interactions with ChatGPT will provide further authentic insight on how to improve the design of assessments and enrich students' learning. Finally, future studies can explore the operationalisation of performance-based assessment using AI interfaces like computer serious games.

References

- [1] 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>.
- [2] 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>.

- [3] Dale, R. (2021) ‘CHATGPT: What’s it good for?’, *Natural Language Engineering*. Cambridge University Press, pp. 113–118. Available at: <https://doi.org/10.1017/S1351324920000601>.
- [4] Elkhatat, A. M. (2023). Evaluating the authenticity of ChatGPT responses: a study on text-matching capabilities. *International Journal for Educational Integrity*, 19(1), 15.
- [5] Mao, R., Chen, G., Zhang, X., Guerin, F., & Cambria, E. (2023). GPTEval: A survey on assessments of ChatGPT and CHATGPT. arXiv preprint arXiv:2308.12488.
- [6] Tamkin, A., Brundage, M., Clark, J., & Ganguli, D. (2021). Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models. <http://arxiv.org/abs/2102.02503>
- [7] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language Models are Few-Shot Learners. 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada. <https://commoncrawl.org/the-data/>
- [8] Langston, J. (2021). Microsoft announces new supercomputer, lays out vision for future AI work. Microsoft News. <https://news.microsoft.com/source/features/ai/openai-azure-supercomputer/> (February 2)
- [9] Hanisch, M., Goldsby, C.M., Fabian, N.E. and Oehmichen, J., 2023. Digital governance: A conceptual framework and research agenda. *Journal of Business Research*, 162, p.113777.
- [10] Jabeur, S.B., Ballouk, H., Arfi, W.B. and Sahut, J.M., 2023. Artificial intelligence applications in fake review detection: Bibliometric analysis and future avenues for research. *Journal of Business Research*, 158, p.113631.
- [11] 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>.
- [12] Rawlasyk, 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>.
- [13] 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>.
- [14] 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>.
- [15] 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>.
- [16] Boud, D. and Falchikov, N. (2006) ‘Aligning assessment with long-term learning’, *Assessment and Evaluation in Higher Education*, 31(4), pp. 399–413. Available at: <https://doi.org/10.1080/02602930600679050>.

- [17] Jirgensons, M. (2015) 'Direct assessment initiatives within a lifelong learning context', in *Procedia Computer Science*. Elsevier B.V., pp. 141–146. Available at: <https://doi.org/10.1016/j.procs.2014.12.019>.
- [18] Belcadhi, L. C. (2016) 'Personalized feedback for self assessment in lifelong learning environments based on semantic web', *Computers in Human Behavior*, 55, pp. 562–570. Available at: <https://doi.org/10.1016/j.chb.2015.07.042>.
- [19] Carless, D., Joughin, G. and Mok, M.M.C. (2006) 'Learning-oriented assessment: Principles and practice', *Assessment and Evaluation in Higher Education*, pp. 395–398. Available at: <https://doi.org/10.1080/02602930600679043>.
- [20] Toro, U., and Joshi, M. (2012). ICT in higher education: Review of literature from the period 2004-2011. *International Journal of Innovation, Management and Technology*, 3(1), 20-23.
- [21] Lockee, B.B. (2021) 'Online education in the post-COVID era', *Nature Electronics*, 4(1), pp. 5–6. Available at: <https://doi.org/10.1038/s41928-020-00534-0>.
- [22] Gallagher, S. and Palmer, J. (2020) 'The Pandemic Pushed Universities Online. The Change Was Long Overdue', *Harvard Business Review*, 29(74).
- [23] 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>.
- [24] Bibauw, S. *et al.* (2022) 'Dialogue systems for language learning: a meta-analysis', *Language Learning & Technology*, 26(1).
- [25] Wollny, S. *et al.* (2021) 'Are We There Yet? - A Systematic Literature Review on Chatbots in Education', *Frontiers in Artificial Intelligence*, 4. Available at: <https://doi.org/10.3389/frai.2021.654924>.
- [26] Tack, A. and Piech, C. (2022) 'The AI Teacher Test: Measuring the Pedagogical Ability of Blender and CHATGPT in Educational Dialogues', *arXiv preprint arXiv:2205.07540* [Preprint]. Available at: <http://arxiv.org/abs/2205.07540>.
- [27] 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>.
- [28] Lagakis, P. and Demetriadis, S. (2021) 'Automated essay scoring: A review of the field', in *Proceedings of the International Conference on Computer, Information, and Telecommunication Systems, CITS 2021*. Available at: <https://doi.org/10.1109/CITS52676.2021.9618476>.
- [29] 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>.
- [30] 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>.
- [31] Craine, K. (2007) 'Managing the Cycle of Change', *Information Management Journal*, 41(5), pp. 44–49.

- [32] Kruse, O. (2006) 'The origins of writing in the disciplines: Traditions of seminar writing and the Humboldtian ideal of the research university', *Written Communication*, 23(3), pp. 331–352. Available at: <https://doi.org/10.1177/0741088306289259>.
- [33] 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>.
- [34] Chuderski, A. (2016) 'Time pressure prevents relational learning', *Learning and Individual Differences*, 49, pp. 361–365. Available at: <https://doi.org/10.1016/j.lindif.2016.07.006>.
- [35] Lynam, S. and Cachia, M. (2018) 'Students' perceptions of the role of assessments at higher education', *Assessment & Evaluation in Higher Education*, 43(2), pp. 223–234. Available at: <https://doi.org/10.1080/02602938.2017.1329928>.
- [36] Whitelock, D. *et al.* (2013) 'OpenEssayist: an automated feedback system that supports university students as they write summative essays', in *The 1st International Conference on Open Learning: Role, Challenges and Aspirations*.
- [37] Newton, P.M. (2018) 'How Common Is Commercial Contract Cheating in Higher Education and Is It Increasing? A Systematic Review', *Frontiers in Education*, 3(67). Available at: <https://doi.org/10.3389/feduc.2018.00067>.
- [38] Elkhatat, A. M., Elsaid, K., and Almeer, S. (2023). Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *International Journal for Educational Integrity*, 19(1), 17.
- [39] Chaka, C. (2023). Detecting AI content in responses generated by ChatGPT, YouChat, and Chatsonic: The case of five AI content detection tools. *Journal of Applied Learning and Teaching*, 6(2).
- [40] 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_clinic_al.45.asp (Accessed: 28 January 2023).
- [41] Li, J., Zhou, X., Wang, Q., (2022). Interventions to reduce loneliness among Chinese older adults: A network meta-analysis of randomized controlled trials and quasi-experimental studies. *Appl Psychol Health Well Being*.
- [42] Sutirman, S., Yuliansah, Y., Isti, R., 2022. The Effectiveness of Electronic Records Information System for Education (ERISE) Applications as a Learning Media of Record Management in the Era of the Industrial Revolution 4.0 698, 1094–1104.
- [43] Wang, L., Wang, C., Yao, X., 2023. Befriended to polarise? The impact of friend identity on review polarisation—A quasi-experiment. In: *Information Systems Journal*. John Wiley and Sons Inc.
- [44] 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>.

- [45] 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>.
- [46] Lindeman, E. (1926) *The meaning of adult education*. 1st edn. New York: New Republic.
- [47] 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>.
- [48] 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>.
- [49] 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>.
- [50] 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>.
- [51] 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>.
- [52] 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>.
- [53] 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>.
- [54] 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>.
- [55] 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.
- [56] 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>.

Appendix 1: Marking rubric for the essay task

Task: Write a 1500-word academic essay on 'The Digital Transformation of the Health Sectors in the Global South' with suggestions on improving digitally enabled healthcare delivery.

Criteria	Outstanding (100-80%)	Excellent (79-70%)	Good (69-60%)	Satisfactory (59-40%)	Poor (0-39%)
Understanding of the Topic	Exceptional understanding with nuanced insights into digital transformation in healthcare in the global south.	Displays an excellent understanding of the topic with comprehensive insights.	Good understanding with some depth and clarity.	Basic understanding shown with general insights.	Poor or no understanding of the topic.
Critical Analysis	Exceptional critical analysis with sophisticated integration of theory and examples.	Provides insightful analysis with well-integrated, relevant theory and examples.	Good level of analysis with relevant theory and examples.	Some analysis, but mostly descriptive or lacking depth.	Little to no critical analysis.
Structure and Organization	Exceptional structure with a compelling and persuasive argument.	Highly logical and coherent structure with a clear and persuasive argument.	Well-organized structure with a clear argument.	Basic structure present but with some disorganization.	Poor structure and organization; argument is hard to follow.
Evidence and Research	Extensive and sophisticated research with excellent integration and citation of sources.	Extensive research with excellent integration and citation of sources.	Good use of research with proper citation of sources.	Some research evident but limited in scope or poorly integrated.	Little to no evidence of research or citation.
Writing Style and Grammar	Exceptional clarity and elegance in writing, error-free.	Exceptionally clear, concise, and error-free writing.	Clear writing with minor errors.	Writing is understandable but with several errors.	Poor writing quality with numerous errors.

Appendix 2: Sample of Grammarly Output

General metrics

4,257	617	26	2 min 28 sec	4 min 44 sec
characters	words	sentences	reading time	speaking time

Score

80

This text scores better than 80% of all texts checked by Grammarly

Writing Issues

37	4	33
Issues left	Critical	Advanced

Writing Issues

4	Correctness	
1	Faulty subject-verb agreement	-----
2	Comma misuse within clauses	-----
1	Determiner use (a/an/the/this, etc.)	-----
6	Clarity	
6	Wordy sentences	-----

Report was generated on Saturday, Dec 23, 2023, 04:05 AM

Page 1 of 6



Report: Untitled

Unique Words

Measures vocabulary diversity by calculating the percentage of words used only once in your document

31%
unique words**Rare Words**

Measures depth of vocabulary by identifying words that are not among the 5,000 most common English words.

35%
rare words**Word Length**

Measures average word length

5.8
characters per word**Sentence Length**

Measures average sentence length

23.7
words per sentence**Appendix 3: Sample of Turnitin originality report (Student 18)**

Candidate 020

ORIGINALITY REPORT

11 %	2 %	0 %	9 %
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Canterbury Christ Church University Student Paper	3 %
2	Submitted to De Montfort University Student Paper	3 %
3	Submitted to Online Education Services Student Paper	3 %
4	www.cms.gov Internet Source	2 %

Exclude quotes On
Exclude bibliography Off

Exclude matches Off

Appendix 4: Sample of Turnitin originality report (Essay 20)

Title: Digital Transformation of Health Sectors in the Global South: Opportunities and Challenges

Introduction

The global landscape of healthcare is undergoing a profound revolution with the advent of digital technologies, offering new avenues for improving healthcare delivery, accessibility, and outcomes. While the developed world has embraced digital transformation in healthcare, the Global South faces unique challenges and opportunities in leveraging technology to enhance its health sectors. This essay explores the digital transformation of health sectors in the Global South, analyzes the current state of digitally enabled healthcare delivery, and provides suggestions for improvement.

Current State of Digital Transformation in the Global South

- Challenges in Infrastructure and Connectivity:** The digital divide is a critical barrier to the widespread implementation of digital health solutions in the Global South. Many regions lack robust information and communication technology (ICT) infrastructure, hindering the seamless integration of digital health tools. In remote areas, poor connectivity further exacerbates the challenges, limiting the reach and effectiveness of digitally enabled healthcare delivery.
- Limited Financial Resources:** Resource constraints in the Global South pose a significant hurdle to adopting advanced digital health technologies. Governments often struggle to allocate funds for building digital health infrastructure, purchasing state-of-the-art equipment, and training healthcare

professionals. As a result, the implementation of digital solutions tends to be slow and uneven across different regions.

3. **Health Information Systems and Interoperability:** The absence of standardized health information systems and interoperability issues present obstacles to the efficient exchange of medical data. In many cases, healthcare providers use disparate systems that do not communicate seamlessly, leading to fragmented patient information and compromising the continuity of care. Addressing these challenges is crucial for the success of digital health initiatives.
4. **Health Literacy and Patient Engagement:** The success of digital health interventions relies heavily on the health literacy of both healthcare providers and patients. In the Global South, limited health literacy, especially in rural areas, can impede the adoption and effective use of digital health tools. Moreover, engaging patients in their own healthcare through digital platforms requires concerted efforts in educating and empowering diverse populations.

Opportunities for Improvement

1. **Investment in ICT Infrastructure:** To overcome the challenges of infrastructure and connectivity, governments and international organizations should prioritize substantial investments in ICT infrastructure. Building robust networks and expanding internet access can create an enabling environment for the deployment of digital health solutions. Public-private partnerships can be instrumental in funding and implementing such initiatives.
2. **Strategic Capacity Building:** Capacity building is crucial for ensuring that healthcare professionals are equipped with the necessary skills to leverage digital technologies effectively. Training programs should be designed to cater to diverse healthcare settings, including urban and rural areas. Collaborative efforts between governments, educational institutions, and technology providers can facilitate the development of a skilled workforce capable of driving the digital transformation of healthcare.
3. **Interoperability Standards and Health Information Exchange:** Establishing interoperability standards and implementing health information exchange mechanisms are imperative for creating a seamless and integrated healthcare ecosystem. Governments and health authorities should work towards adopting standardized protocols that enable the secure and efficient sharing of health data across different platforms. International collaboration can contribute to the development and adoption of global interoperability standards.
4. **Community Engagement and Health Literacy Programs:** Promoting health literacy and community engagement is essential for the successful adoption of digital health solutions. Public awareness campaigns, community workshops, and educational initiatives can empower individuals to actively participate in their healthcare. Tailoring communication strategies to local languages and cultural contexts is crucial for overcoming barriers related to health literacy in diverse populations.
5. **Telemedicine and Remote Patient Monitoring:** Leveraging telemedicine and remote patient monitoring can bridge the geographical gap between healthcare providers and patients, particularly in remote areas. Governments should invest in telecommunication infrastructure to support telemedicine services and facilitate remote consultations. This approach can enhance access to healthcare services and improve health outcomes, especially for populations with limited physical access to healthcare facilities.
6. **Mobile Health (mHealth) Solutions:** The widespread use of mobile phones in the Global South presents a unique opportunity for implementing mobile health solutions. Developing user-friendly health apps, SMS-based interventions, and mobile platforms for health information dissemination can empower individuals to take charge of their health. Additionally, governments can explore partnerships with mobile network operators to expand the reach of mHealth initiatives.

Conclusion

The digital transformation of health sectors in the Global South holds immense potential for improving healthcare delivery, accessibility, and outcomes. While challenges such as infrastructure limitations, financial constraints, and health literacy barriers persist, strategic interventions can pave the way for meaningful progress. Governments, international organizations, healthcare providers, and technology developers must collaborate to address these challenges and capitalize on the opportunities presented by digital transformation.

By investing in ICT infrastructure, implementing interoperability standards, and prioritizing capacity building, the Global South can create a foundation for sustainable digital health initiatives. Community engagement and targeted health literacy programs are essential components of this transformation, ensuring that individuals are active participants in their healthcare journey. Embracing telemedicine, remote patient monitoring, and mobile health solutions can further extend the reach of healthcare services, especially to underserved populations.

In conclusion, the digital transformation of health sectors in the Global South is a multifaceted endeavor that requires a comprehensive and collaborative approach. By addressing the unique challenges and capitalizing on the opportunities, the Global South can harness the power of digital technologies to build resilient and inclusive healthcare systems, ultimately improving the well-being of its populations.

Highlights

- ChatGPT generates high quality, original essays from unique user accounts
- ChatGPT is unable to generate multiple original contents from the same user account
- ChatGPT struggles with referencing and task completion in response to prompts
- AI can mediate an integrated assessment framework that combines knowledge testing with competence assessment and performance evaluation

Journal Pre-proof

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Journal Pre-proof