Susceptibility of Assessment Types to AI-Generated Content: a Quasi-Experimental Pilot Study in Digital Health and Health Information Management Education

Tafheem Ahmad Wani, Michael Liem, Natasha Prasad, Kerin Robinson, Abbey Nexhip, Melanie Tassos, Stephanie Gjorgioski, Urooj Raza Khan, James Boyd, Merilyn Riley

Submitted to: JMIR Medical Education
on: August 26, 2025

Table of Contents

Susceptibility of Assessment Types to AI-Generated Content: a Quasi-Experimental Pilot Study in Digital Health and Health Information Management Education

Tafheem Ahmad Wani[1]; Michael Liem[1]; Natasha Prasad[1]; Kerin Robinson[1]; Abbey Nexhip[1]; Melanie Tassos[1]; Stephanie Gjorgioski[1]; Urooj Raza Khan[1]; James Boyd[1]; Merilyn Riley[1]

[1] La Trobe University Melbourne AU

Corresponding Author:
Tafheem Ahmad Wani

La Trobe University
Health Sciences 2, La Trobe University, Bundoora
Melbourne
AU

Abstract

Background: Generative artificial intelligence (GenAI) tools, such as ChatGPT, are reshaping higher education and prompting urgent discussions about academic integrity. In Digital Health and Health Information Management (DIGHIM) programs, where assessment tasks often require a combination of technical proficiency, contextual reasoning, and professional judgment, the integration of GenAI presents unique opportunities and risks. These programs train graduates to work at the intersection of health, data, and technology, making it essential to understand how AI performs across the diverse assessment formats that reflect real-world professional competencies.

Objective: The pilot study aimed to evaluate ChatGPT's performance across diverse assessment types in DIGHIM education by examining how task complexity influences AI-generated output quality, and develop recommendations for ethical and effective AI integration in assessments.

Methods: A pilot quasi-experimental design compared ChatGPT-generated responses with de-identified student submissions across five assessment types: digital health solution design, business case analysis, reflective assessment, SQL health database programming, and a health classification quiz. For each task, multiple AI submissions were produced using different prompting strategies, including rubric integration and the use of ChatGPT-4.0 and o1 Preview. Blinded academic markers evaluated all submissions against standard rubrics, and descriptive statistics were used to compare performance.

Results: ChatGPT's performance varied considerably across assessment types. It achieved its highest accuracy in objective, rule-based tasks such as multiple-choice quiz items in health classification (mean 87.5%) and produced well-structured, coherent responses for reflective assessments (mean 69.4%), though these often-lacked personalisation and nuanced industry context. In descriptive analytical tasks, such as digital health business cases and solution designs, ChatGPT produced logically structured work with reasonable use of evidence but failed to provide deep contextualisation, domain-specific insights, or visual elements expected in DIGHIM practice. Technical assessments revealed the greatest limitations: SQL programming tasks averaged 42.3%, with persistent schema errors, incomplete queries, and weak interpretation of health data outputs, while scenario-based clinical coding scored just 7.1%, reflecting a lack of precision in applying ICD-10-AM rules and coding conventions. Structured prompting and rubric integration improved results, particularly in descriptive and reflective tasks (up to 80%), but the advanced ChatGPT o1 Preview model did not consistently outperform earlier versions.

Conclusions: While ChatGPT demonstrates strong capability in structured, rule-based, and reflective tasks, it remains limited in technical accuracy, contextual reasoning, and application to Digital Health and Health Information Management contexts. To preserve academic integrity and ensure graduates are workforce-ready, assessment designs should emphasise critical thinking, ethical reasoning, and scenario-based problem-solving that reflect real-world DIGHIM practice. Integrating AI as a tool for critique, refinement, and validation, rather than as a replacement for student work can help educators prepare students for responsible AI use in digital health and health information management professions.

(JMIR Preprints 26/08/2025:82988)
DOI: https://doi.org/10.2196/preprints.82988

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http
No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <a href="https:/

Original Manuscript

Susceptibility of Assessment Types to AI-Generated Content: a Quasi-Experimental Pilot Study in Digital Health and Health Information Management Education

Abstract

Background: Generative artificial intelligence (GenAI) tools, such as ChatGPT, are reshaping higher education and prompting urgent discussions about academic integrity. In Digital Health and Health Information Management (DIGHIM) programs, where assessment tasks often require a combination of technical proficiency, contextual reasoning, and professional judgment, the integration of GenAI presents unique opportunities and risks. These programs train graduates to work at the intersection of health, data, and technology, making it essential to understand how AI performs across the diverse assessment formats that reflect real-world professional competencies.
Objective: The pilot study aimed to evaluate ChatGPT's performance across diverse assessment types in DIGHIM education by examining how task complexity influences AI-generated output quality, and develop recommendations for ethical and effective AI integration in assessments.
Methods: A pilot quasi-experimental design compared ChatGPT-generated responses with de-identified student submissions across five assessment types: digital health solution design, business case analysis, reflective assessment, SQL health database programming, and a health classification quiz. For each task, multiple AI submissions were produced using different prompting strategies, including rubric integration and the use of ChatGPT-4.0 and o1 Preview. Blinded academic markers evaluated all submissions against standard rubrics, and descriptive statistics were used to compare performance.
Results: ChatGPT's performance varied considerably across assessment types. It achieved its highest accuracy in objective, rule-based tasks such as multiple-choice quiz items in health classification (mean 87.5%) and produced well-structured, coherent responses for reflective assessments (mean 69.4%), though these often-lacked personalisation and nuanced industry context. In descriptive analytical tasks, such as digital health business cases and solution designs, ChatGPT produced logically structured work with reasonable use of evidence but failed to provide deep contextualisation, domain-specific insights, or visual elements expected in DIGHIM practice. Technical assessments revealed the greatest limitations: SQL programming tasks averaged 42.3%, with persistent schema errors, incomplete queries, and weak interpretation of health data outputs, while scenario-based clinical coding scored just 7.1%, reflecting a lack of precision in applying ICD-10-AM rules and coding conventions. Structured prompting and rubric integration improved results, particularly in descriptive and reflective tasks (up to 80%), but the advanced ChatGPT o1 Preview model did not consistently outperform earlier versions.
Conclusion: While ChatGPT demonstrates strong capability in structured, rule-based, and reflective tasks, it remains limited in technical accuracy, contextual reasoning, and application to Digital Health and Health Information Management contexts. To preserve academic integrity and ensure graduates are workforce-ready, assessment designs should emphasise critical thinking, ethical reasoning, and scenario-based problem-solving that reflect real-world DIGHIM practice. Integrating AI as a tool for critique, refinement, and validation, rather than as a replacement for student work can help educators prepare students for responsible AI use in digital health and health information management professions.

Keywords: generative artificial intelligence; academic integrity; assessment design; digital health education; health information management; ChatGPT performance; quasi-experimental study

Introduction

Artificial intelligence (AI) has emerged as one of the most transformative technologies of the 21st century, redefining the way humans interact with machines [1]. The International Organization for Standardisation (ISO) defines AI as “a technical and scientific field devoted to the engineered system that generates outputs such as content, forecasts, recommendations or decisions for a given set of human-defined objectives” (ISO/International Electrotechnical Commission) [2]. Among the many branches of AI, generative AI (GenAI) stands out as a groundbreaking development. GenAI encompasses advanced systems that generate human-like media, such as text, images, videos, and code in response to user prompts [3]. At the forefront of this revolution are large language models (LLMs) such as ChatGPT, Microsoft Co-Pilot and Gemini, which have rapidly expanded the potential for AI applications across fields such as healthcare, business, and engineering [4–6]. Powered by deep learning and advanced algorithms, LLMs excel in language-related tasks, such as generating text, answering questions, and understanding context to produce human-like responses [7,8], in a variety of languages [9]. Depending on the user’s context, it can also adapt its language style, tone, and formality to suit a range of communicative purposes [9,10]. With extensive training data and billions of parameters, these models are continually refined through reinforcement learning and human feedback, thereby enabling increasingly sophisticated performance [11].

Due to their ability to process and fine-tune increasingly complex questions, LLMs, in particular ChatGPT, have quickly gained the attention of academics and students in higher education [8,12–14]. Several researchers have explored how academics utilise ChatGPT in assignments, assessments, and examination design, including its potential to automate marking practices [14–17]. These applications can improve the timeliness of feedback and minimise human grading errors [14,18,19]. Academics can also benefit from ChatGPT use in curriculum design [17]. ChatGPT can assist students in refining their written language by suggesting corrections to grammatical and syntactical structures, identifying errors, and providing vocabulary enhancements. It can also be used to generate ideas and research questions as a starting point for further inquiry [20].

Despite the benefits, the integration of GenAI technologies into higher education can present significant ethical challenges [21]. These include ChatGPT’s use of inaccurate content (including fictitious reference material) [22] concerns regarding human-teacher replacement [11], and negative repercussions on students’ critical thinking and problem-solving skills [23]. Most significantly, the use of ChatGPT can pose a threat to academic integrity and ethics [24]. A recent 2024 scoping review has highlighted that traditional assessment methods do not operate effectively in GenAI-facilitated learning environments, prompting the need for innovative and refocused assessment designs that foster career-driven competencies and lifelong learning skills [25].
The ethical framework surrounding students’ use of AI in educational assessments primarily involves transparency, accountability, and reliability. Transparency requires that students acknowledge the use of AI appropriately in their submissions. Students must also be accountable for the content they submit, as “outputs of AI tools can include biased, inaccurate, or incorrect content that users should be aware of” [26]. In some instances, generative AI can fabricate information, which raises the question of the reliability of the outputs [7].

The accessibility of AI tools and the difficulty in detecting AI-generated content [7,16] may tempt students to engage in academic misconduct. This raises concerns about assessment fairness, equity, and the credibility of academic credentials earned through digital platforms. Educators, therefore, face challenges in maintaining the integrity of assessments and ensuring educational quality [16,27].

La Trobe University, Australia, offers a wide range of health programs. The Digital Health and

Information Management (DIGHIM) programs provide formal education for these disciplines, aiming to develop students' technical, professional and interpersonal skills [28,29]. These programs emphasise discipline-specific and generic competencies, which are inclusive of problem-solving, critical thinking, and ethical decision-making that are typically assessed through academic essays, reports, multiple-choice quizzes, presentations, case studies, programming exercises, and practical simulations [30,31]. The proliferation of AI-generated content threatens the validity and reliability of assessments in most of these areas, potentially compromising the effectiveness of evaluating students' comprehension and competence. [16].

Previous research on the impact of GenAI tools in higher education has primarily focused on specific types of assessments, such as multiple-choice questions or essay evaluations, mainly within the medical [32,33] and business fields [4]. For example, Chaudhry et al. [16] undertook evaluation of AI-generated assessments in a Bachelor of Business Administration, including case analysis, empirical study report, self-reflection group work, and calculation-based assessments. These findings do not necessarily translate well to the more specialised fields of Digital Health (DH) and Health Information Management (HIM), each of which requires a distinct combination of technical, analytical, and decision-making skills. DIGHIM programs, for instance, demand expertise in areas such as health classification, epidemiology and biostatistics, digital health solution design, health data governance, and interoperability [28,29]. These specialist areas of study necessitate a rigorous and diverse set of assessment methods, incorporating scenario-based problem-solving, data interpretation, and system implementation tasks that test both theoretical knowledge and practical application. Given the interdisciplinary nature of digital health, which aims to bridge clinical practice, data science, and information technology, there is a need for a separate evaluation of how GenAI interacts with these unique assessments.

Aims

This pilot study aimed to explore how the complexity of different assessment tasks influences the quality and reliability of AI-generated content in DIGHIM courses. The overarching goal was structured into three key objectives:

1. To generate preliminary recommendations for the responsible and ethical use of generative AI in student learning, with a focus on ensuring assessments foster critical thinking and technical proficiency.
2. To trial approaches that may equip educators with strategies for designing academically rigorous assessments that support responsible AI engagement.
3. To examine how AI-authentic assessments can contribute to preparing students for the evolving demands of digital health and health information management professions, particularly in fostering professional competencies.

Methods

Study design

A quasi-experimental design was piloted to evaluate ChatGPT's performance on a range of assessment types, comparing its results with submitted assessments of past students. This exploratory approach was intended to provide preliminary insights and inform the design of larger-scale studies. While similar methods have been applied in ChatGPT response evaluation within the education sector, this study pilots the approach in the context of Digital Health and Health Information Management, where it has not yet been examined[11,16,33].

Assessment materials and Markers (Participants)

Assessments from five DIGHIM subjects were purposively selected because of specific assessment characteristics, to maximise diversity and breadth in the type and nature of assessments chosen. The assessment types included i) an online quiz/examination, ii) SQL programming, iii) reflective practice assessment, iv) business case proposal report, and v) digital health solution design (see Table 1). Past students' de-identified assessments, originally submitted for academic credit, were re-evaluated for research purposes as a benchmark against which the ChatGPT-generated responses were compared. Markers, blinded to the origin of each assessment, were responsible for evaluating both the ChatGPT-generated assessments and the past students' assessments.

Sample

Control group: Subject coordinators selected a sample of student assessments (three per each assessment type) completed between 1 July 2023 and 30 June 2024 to form the control group. To ensure representativeness, assessments were chosen across three grade bands: high performance (80% and above), medium performance (70–79%), and low performance (50–59%). This resulted in the inclusion of a total of 15 student assessments across the five subjects and five assessment types. All assessments were de-identified prior to analysis to maintain student anonymity.

Experimental group: ChatGPT was used to complete the same assessment tasks as were selected for the control group. Two to four AI-generated assessments were created for each assessment type (Table 1, $n = 18$) to reflect variations in how a typical student might approach the task. The criteria for these variations and their alignment with student input patterns are detailed below in the experiment section.

In total, 33 assessments, across the five assessment types, were selected for the study: 18 were AI generated (experimental group), and 15 were past students' assessments (control group).

Table 1: Summary of assessment types/sample chosen for the study

Assessment Type	Level	Discipline	Assessment Description	Experimental (AI) Group: Assessments Marked	Control (Student) Assessments Marked
Digital Health Solution Design Approach	Masters	Digital Health	Propose and justify a structured approach to addressing a specific healthcare challenge through a digital health solution, incorporating design principles, implementation frameworks, and the consideration of barriers and enablers.	4	3
Digital Health	Masters	Digital Health	Develop a proposal identifying gaps in a	4	3

Business Case Proposal Report/Case Study			digital health system, analysing barriers, and recommending evidence-based, innovative solutions.		
Health Information Management Reflective Assessment	4th Year Undergraduate	Health Information Management	Reflection on health information management students' professional practice (Work-Integrated Learning) placement experience and learning	4	3
SQL/ Programming	2nd Year Undergraduate	Health Information Management	Use of SQL to query and analyse a health database, generating reports and extracting relevant data	4	3
Health Classification Online Quiz Examination	1st Year Undergraduate	Health Information Management	Online examination comprising long, short, and objective (MCQ, True/False and one word answer) questions, assessing knowledge of health classification systems (ICD-10-AM) and clinical coding principles	2	3

Experiment

An independent research assistant (RA) followed a structured process, established by the researchers, to gather ChatGPT assessment responses for the chosen sample of assessments (Figure 1).

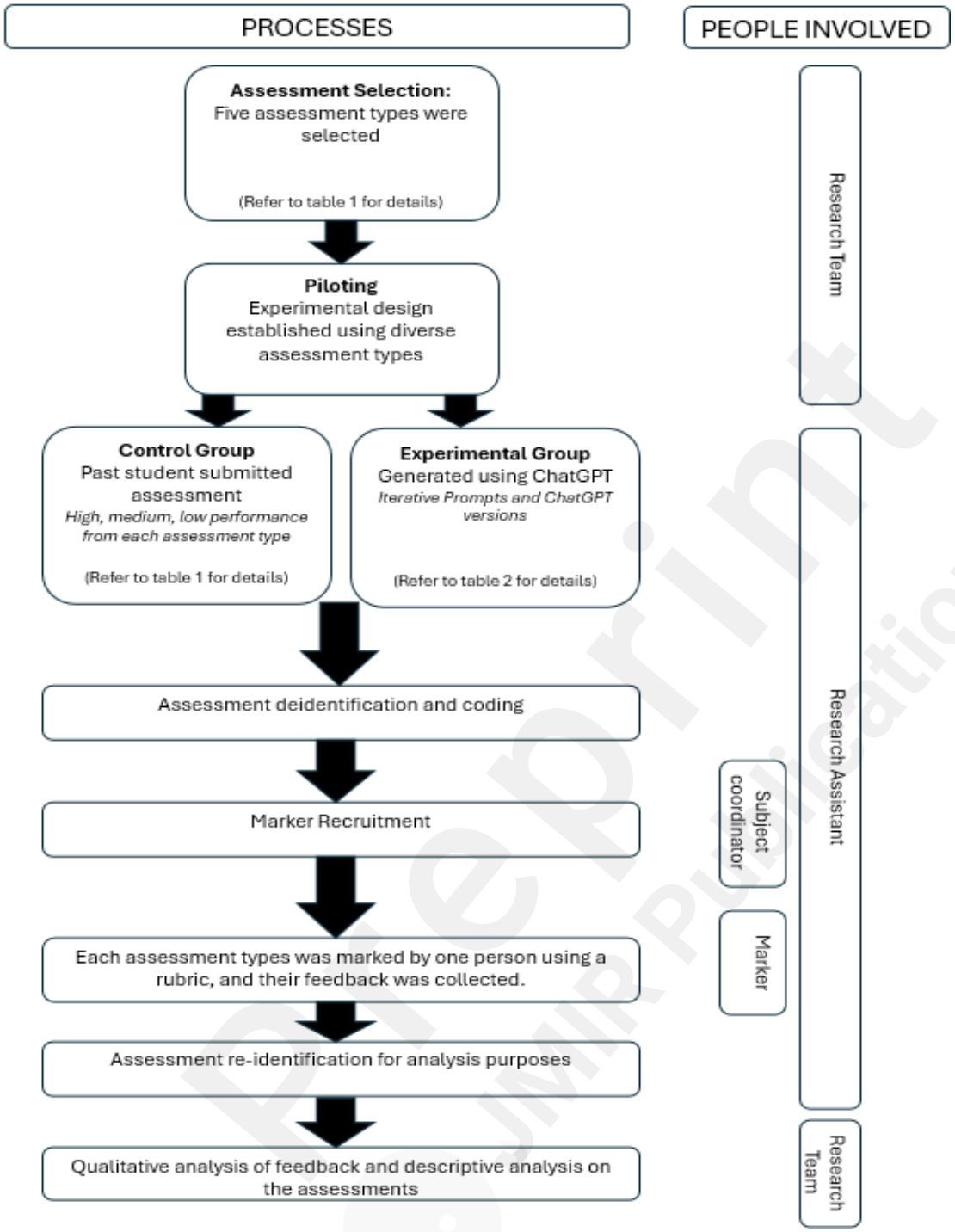


Figure 1: Overview of methodology

a) *Generating ChatGPT responses*

To ensure reliability, completeness, and alignment with academic expectations, multiple submission versions of assessment instructions were developed for each assessment type for submission to ChatGPT and its versions (Table 2).

Table 2: Description of AI-generated submission versions and prompting strategies

Assessment Type	Version Name		Description
Subjective and analytical assessments (Reflective Assessment, Business Case Proposal/Case Study, Digital Health Solution)	ChatGPT (submission v1)	4.0	Full assessment instructions entered into ChatGPT 4.0 for a single-step response generation.
	ChatGPT (submission v2)	4.0	Assessment broken into parts as per instructions; ChatGPT 4.0 generated each part separately and compiled.
	ChatGPT (submission v3)	4.0	Same as Version 2, with the rubric provided to guide response generation for better alignment with criteria.
	ChatGPT (submission v4)	o1	Same as Version 3, but generated using ChatGPT o1 Preview to evaluate differences in performance.
SQL/ Programming Assessment	ChatGPT (submission v1)	4.0	Query instructions provided without schema for response generation.
	ChatGPT (submission v2)	4.0	Query instructions with schema provided to enhance contextual accuracy.
	ChatGPT (submission v3)	4.0	Schema and rubric provided to ensure responses aligned with evaluation criteria.
	ChatGPT (submission v4)	o1	Same as Version 3, but generated using ChatGPT o1 preview for comparative analysis.
Health Classification Online Quiz Examination (Scenario based and objective questions)	ChatGPT (submission v1)	4.0	Objective questions generated using ChatGPT 4.0.
	ChatGPT (submission v2)	o1	Objective questions generated using ChatGPT o1 Preview to compare differences in outputs.

It was expected that full-instruction (Version 1) and step-by-step approaches (Versions 2 and 3) would allow for a nuanced analysis of AI-generated outputs, with versions 2 and 3 specifically designed to assess how incremental guidance impacted response quality. Furthermore, inclusion of rubrics (Version 3) was expected to further improve contextual alignment. The inclusion of ChatGPT o1 Preview (Version 4) used the ChatGPT o1 model, which is designed for advanced reasoning, and provided insights into GenAI's evolving capabilities. For SQL tasks, submission version (Versions 2 and 3) included the database schema and rubric for enhanced query accuracy. In contrast, for objective-style questions in the Health Classification assessment, only two submission versions were needed due to the structured format of the questions and the straightforward nature of the required responses.

b) *Blinded review preparation:*

The research assistant formatted ChatGPT-generated responses to resemble student submissions. These AI-generated responses were then mixed with real, de-identified student assessments to maintain blinding. To ensure an unbiased review process, all assessments were anonymised and assigned coded identifiers.

c) *Assessment allocation:*
Based on their expertise in the subject, each subject coordinator suggested potential markers to the RA, to assess and mark both ChatGPT responses and student assessments. The RA contacted potential markers individually via email to seek their participation and to shortlist one marker per assessment type. Markers were provided with Participant Information Consent Forms (PICF). To maintain impartiality, the nominated markers did not include the academics who originally marked the student assessments. Assessments were randomly assigned to the markers by the RA, acting as an independent coordinator to ensure a fair distribution and prevent any conflicts of interest.

d) *Marking process*:
Markers followed a standardised marking rubric to evaluate the assessments. To ensure consistency in grading, clear instructions were provided by the principal investigator and communicated by the independent RA. Additionally, support was available throughout the process to address any questions or uncertainties. In addition to numerical scores, markers provided written feedback aligned with the rubric criteria, along with overall free-form comments to justify their grading and offer insights into the quality of the work. To further ensure reliability, if a discrepancy of more than 10 marks was identified in any student assessment (between the original mark and remark), all assessments within that assessment type were independently reviewed by a second subject expert. Any differences were then discussed, and a consensus was reached to finalise the grades.

e) *Data tracking and management:*
The RA maintained a Microsoft Excel spreadsheet to track the marking process, including details of assessments sent to staff members and their feedback.

f) *Testing and review:*
As part of the experimental process, an initial round of testing was undertaken to refine the methodology and confirm the feasibility of the planned research. This involved a smaller sample of assessments across various types (Table 2) to trial the generation of AI responses, the collection and de-identification of student work, and the blinded review process (steps b–e). Insights from this stage were used to make necessary adjustments prior to the main experiment. The assessments used during this testing phase were excluded from the final data analysis.

Synthesis and analysis

Marks and qualitative feedback provided by academic markers were compiled for each assessment type. Descriptive statistics (means, percentages, and score ranges) were calculated to compare ChatGPT outputs with the highest-graded student submissions. Performance was examined both within individual assessment types (Tables 3–7) and across assessment categories (descriptive, reflective, technical, and objective tasks).

Comparative trend analysis was undertaken to evaluate progression across ChatGPT versions (V1–V4), focusing on improvements associated with structured prompting, schema inclusion, and rubric integration. In addition, rubric-level synthesis was conducted to identify recurring strengths and weaknesses.

Feedback comments were thematically coded to detect common indicators of AI-generated content, such as formulaic phrasing, generic recommendations, and fabricated references. These were contrasted with feedback patterns for top student submissions to highlight areas where AI outputs diverged most significantly from human-authored work.

Results were further synthesised into cross-assessment comparisons, with performance grouped into broader task categories (objective, reflective, descriptive analytical, scenario-based analytical, communication/referencing, and programming) to identify where AI demonstrated relative strengths versus persistent limitations (Figures 2–3)

Generative AI tools used

The study used ChatGPT 4.0 (free version/limited) and ChatGPT o1 preview (pro/paid version) for the experimentation. The researchers opted out of ChatGPT's AI training model using an available feature, ensuring that the submitted assessment guidelines and instructions were not used to further train the AI. All necessary privacy settings were enabled to maintain data confidentiality throughout the research process.

Ethics approval

The research study was approved by La Trobe University's Human Research Ethics Committee (Application No. HEC24286).

Results

Performance by assessment type

This section provides a detailed breakdown of ChatGPT's performance across each assessment type, with comparisons to the highest-graded student assessments. Comprehensive data analysis and full qualitative feedback for all assessments are provided in multimedia appendix 1: Detailed performance analysis by assessment type.

Assessment Type 1: Design of Digital Health Solution

This assessment required students to design a digital health solution addressing a specific healthcare challenge, incorporating implementation frameworks and a critical discussion of barriers and enablers. Submissions were evaluated using eight rubric criteria, including problem framing, justification of design, contextual relevance, and professional communication.

AI-generated responses showed progressive improvements with structured prompting and rubric integration, most notably in ChatGPT 4.0 V3, which scored 69% compared to 56% for ChatGPT 4.0 V1 (Table 3). Overall, even with these refinements, AI outputs lacked the depth, specificity, and contextual alignment observed in student submissions. Notably, the use of the more advanced ChatGPT o1 model (V4) did not improve performance over ChatGPT 4.0 V3, scoring only 63%, and continued to exhibit generic recommendations and limited tailoring to the scenario. In contrast, the top-performing student submission (81%) demonstrated nuanced justification, visual clarity, and strong alignment with project objectives.

Table 3: Comparison of outcomes from ChatGPT submission versions for Digital Health Solution Assessment

Submission Version Name/Number	Score/ 100	Key positive points noted by marker	Key points for improvement noted by marker
4.0 V1	56	Design stages moderately justified	Vague barrier discussion; unclear language; misplaced citations

4.0 V2	60	Improved clarity and referencing (than v1); better articulation of barriers	Generic roadmap activities; limited alignment with goals
4.0 V3	69	Design approach supported with clear evidence; better understanding of challenges; strong structure; clearer communication; improved referencing	Activities remained broad; lacked visual aids and specificity
o1 V4	63	Consistent framework discussion; professional tone	Roadmap remained generic; limited contextual alignment

Assessment Type 2: Digital Health Business Proposal/Case Study

This assessment required students to develop a comprehensive business proposal that identified gaps in an existing digital health system and recommended innovative technological solutions. The task emphasised evidence-based justification, analysis of implementation barriers, and a professional format capable of demonstrating the proposed solution's value and impact on healthcare delivery.

ChatGPT-generated submissions demonstrated a basic understanding of the task and performed consistently across versions (Table 4). ChatGPT 4.0 V2 scored the highest (68%), showing improved referencing and clearer structure. However, despite incremental refinements, all AI outputs were critiqued for their generic content, lack of tailored insights, and absence of visual elements such as patient journey maps or system flow diagrams. Notably, the more advanced ChatGPT o1 model (V4) did not improve performance (score: 64%) and continued to present directive rather than strategic framing. In contrast, the top-performing student submission (score: 82%) demonstrated personalised problem-solution alignment, stronger analytical depth, and enhanced clarity through well-integrated visual aids.

Table 4: Comparison of outcomes from ChatGPT submission versions - Digital Health Business Proposal/Case Study assessment

Submission Version Name/Number	**Score/ 100**	**Key positive points noted by marker**	**Key points for improvement noted by marker**
ChatGPT 4.0 V1	62	Basic identification of system gaps; logical structure	Weak justification of solutions; insufficient references; vague insights
ChatGPT 4.0 V2	68	Clearer articulation of barriers; improved referencing and layout	Lacked personalisation; limited strategic alignment with healthcare needs
ChatGPT 4.0 V3	67	Stronger evidence-based rationale; professional tone	Solutions not well integrated with patient needs; absence of visual aids such as journey maps
ChatGPT o1 V4	64	Consistently identified	Lacked strategic framing;

	issues and proposed digital interventions	weak linkage between problems and proposed actions

Assessment Type 3: Health Information Management - Reflective Assessment

This assessment required students to critically reflect on their placement experience within a Health Information Management context, integrating evidence from empirical studies to evaluate professional competencies. Students were expected to identify their strengths and weaknesses and to set actionable goals for growth while demonstrating professional communication.

ChatGPT-generated responses showed a wide range in performance (Table 5). ChatGPT 4.0 V3 achieved the highest AI score (80%), presenting a well-structured report with detailed reflections and systematic use of empirical evidence. However, it exceeded the word limit and lacked visual enhancements, while some goals remained broad and misaligned with competencies expected in the health industry. Earlier versions showed key weaknesses: V1 (53%) was concise but underdeveloped and lacked depth in reflection and linkage to placement experience; V2 (70%) improved on structure and relevance but included fabricated references and overly clinical goals. The most advanced version, ChatGPT o1 V4, performed poorly (55%)—meeting word limits but offering shallow insights and repetitive, vague reflections.

While ChatGPT demonstrated fluency, structure, and effective use of reflective frameworks, it fell short in personalisation, contextual relevance, and depth, elements that distinguished top student submissions, which provided clear placement-specific insights, nuanced analysis, and realistic strategies for professional development.

Table 5: Comparison of Outcomes from ChatGPT Submission Versions – HIM Reflective Assessment

Submission Version Name/Number	Score/ 100	Key positive points noted by marker	Key points for improvement noted by marker
ChatGPT 4.0 V1	53	Structured format; identified gaps in reflective practice	Omitted details; weak placement linkage; lacked depth and clarity
ChatGPT 4.0 V2	70	Strong organisation; good competency linkage, visual structure	Overly clinical focus; exceeded word count; included fabricated reference
ChatGPT 4.0 V3	80	Comprehensive, well-referenced; detailed placement reflection	Wordy; lacked visuals; some goals too broad for HIM context
ChatGPT o1 V4	55	Concise and grammatically sound; some relevant evidence	Shallow reflection; repetitive content; weak critical analysis

Assessment Type 4: Health Database SQL Programming

This assessment evaluated students' ability to apply SQL programming skills to analyse a health dataset. Students were required to construct accurate queries, extract relevant health-related information, and present meaningful data insights to support decision-making in healthcare settings. The task assessed identification of health information needs, correct selection of tables and variables, query accuracy, and clarity in result interpretation.

AI-generated responses demonstrated only modest improvement across iterations, with schema and rubric integration resulting in higher scores (Table 6). ChatGPT 4.0 V1, generated without schema input, performed poorly (17%), with all queries based on incorrect table and column names. Schema inclusion in V2 improved structural accuracy (47%) but errors in data grouping and missing outputs persisted. V3 (49%) refined query logic slightly, though misspellings and omission of key fields remained problematic. The most advanced model, ChatGPT o1 V4, achieved the highest AI score (56%), correctly executing about half of the queries and demonstrating improved adherence to SQL conventions—but still suffered from inaccuracies in age group breakdowns, incorrect counts, and missing contextual details like diagnosis descriptions.

In contrast, the top student submission showed precise application of SQL logic, full query completion, and context-aware interpretation of health trends. It demonstrated superior attention to schema structure, data accuracy, and professional communication in presenting findings—areas where ChatGPT consistently underperformed.

Table 6: Comparison of outcomes from ChatGPT submission versions – Health Database SQL Programming

Submission Version Name/Number	Score/ 100	Key positive points noted by marker	Key points for improvement noted by marker
ChatGPT 4.0 V1	17	Correct use of SQL structure; basic understanding of syntax	All table/column names incorrect; no valid outputs; missing queries
ChatGPT 4.0 V2	47	Some correct queries; schema improved structural logic	Wrong grouping; spelling errors; partial outputs; inconsistent results
ChatGPT 4.0 V3	49	Better rubric alignment; clearer structure	Inaccurate grouping; incomplete queries; diagnosis data omitted
ChatGPT o1 V4	56	Improved accuracy; half the queries returned correct results	Ongoing errors in counts and terminology; limited interpretation

Assessment Type 5: Health Information Management Health Classification Online Quiz Examination

This quiz assessed students' application of ICD-10-AM health classification standards through scenario-based coding, line coding, and objective questions.

AI responses (ChatGPT 4.0 V1 and Preview V2) performed well in the objective section (87.5%) of the quiz but struggled with the scenario-based coding tasks, scoring 32.4% overall, over 57% lower than the top student score (Table 7). Key issues included incorrect block numbers, inaccurate sequencing, and vague or missing code justifications. In contrast, the top student demonstrated precision and contextual understanding aligned with national coding standards. These results reinforce that while AI handles structured recall effectively, it lacks the nuanced reasoning required for applied clinical coding.

Table 7: Comparison of Outcomes from ChatGPT Submission Versions – Health Classification Quiz

Submission Version Name/Number	Score/ 100	Key positive points noted by marker	Key points for improvement noted by marker
ChatGPT 4.0 V1	32.4	Strong performance in objective section (14/16 correct)	Incorrect block numbers; poor sequencing; inaccurate tabular usage
ChatGPT 4.0 Preview V2	32.4	Consistent objective accuracy; good understanding of conventions	Weak in scenario coding; missing procedural codes; vague codes/justifications

**Only two versions were tested, as rubric- and breakdown-based versions were not applicable to this quiz format.*

Cumulative performance comparison across assessment types and tasks

Comparative analysis across assessment types

An overall performance comparison across different versions of ChatGPT responses, based on average scores, revealed noticeable variations in performance across assessments (Figure 2).

Descriptive (Digital Health case study assessment average: 65.3%, Digital Health solution design assessment average: 62%) and reflective assessments (HIM reflective assessment average: 64.5%) consistently achieved higher average scores in Chat GPT-generated responses compared to technical tasks like SQL programming (assessment average: 42.3%) and health classification quiz (assessment average: 32.4%).

Conversely, in technical assessments such as SQL programming and health classification, ChatGPT performed poorly. Lower averages in these tasks reflected the complexity of accurately interpreting technical prompts, adhering to database structures, and executing precise scripts. Progression from V1 to V4 in technical tasks demonstrated the importance of context, schema inclusion, and rubrics, which helped improve performance over iterations. However, even the best-performing versions struggled to match the precision and depth required for these tasks, underlining limitations in handling computational assessments.

Overall, V3 emerged as the most consistently strong version across assessment types, benefiting from rubric-driven structuring and better alignment with task criteria. V2 showed notable success in descriptive tasks due to its structured approach, while V4 demonstrated improvements in technical assessments through contextual and schema support. In contrast, V1 consistently underperformed

across most assessments, reflecting its reliance on single-step, unguided generation, which often lacked depth, accuracy, and alignment with task expectations.

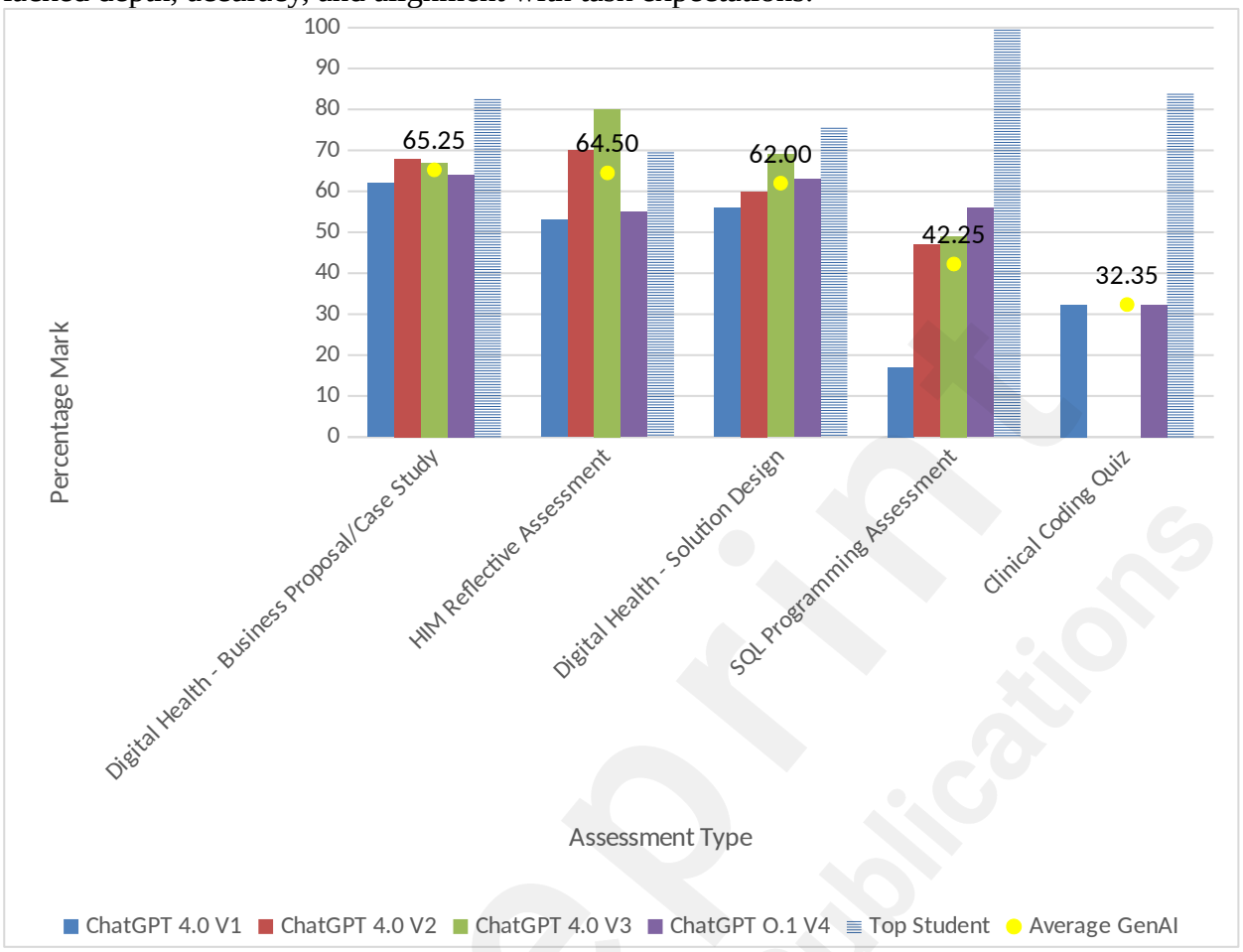


Figure 2: Overall performance comparison across assessment types

Comparative Analysis Across Task Types

To evaluate ChatGPT's performance, similar tasks from multiple assessments were grouped into broader categories: objective, reflective, descriptive analytical, scenario-based analytical, communication/referencing, and programming. ChatGPT performed best on objective tasks (average 87.5%), showing strong accuracy in factual, rule-based questions. Reflective tasks followed (69.4%), with later versions demonstrating improved reasoning. Descriptive analytical tasks had moderate performance (63.0%), though gains plateaued in newer versions. Communication and referencing tasks averaged 60.9%, with a clearer structure in later iterations. Programming tasks improved over time (42.3%) but continued to face challenges in accuracy and contextual understanding. The lowest performance was in complex scenario-based health classification tasks (7.1%), reflecting current limitations in contextual and analytical reasoning (Figure3).

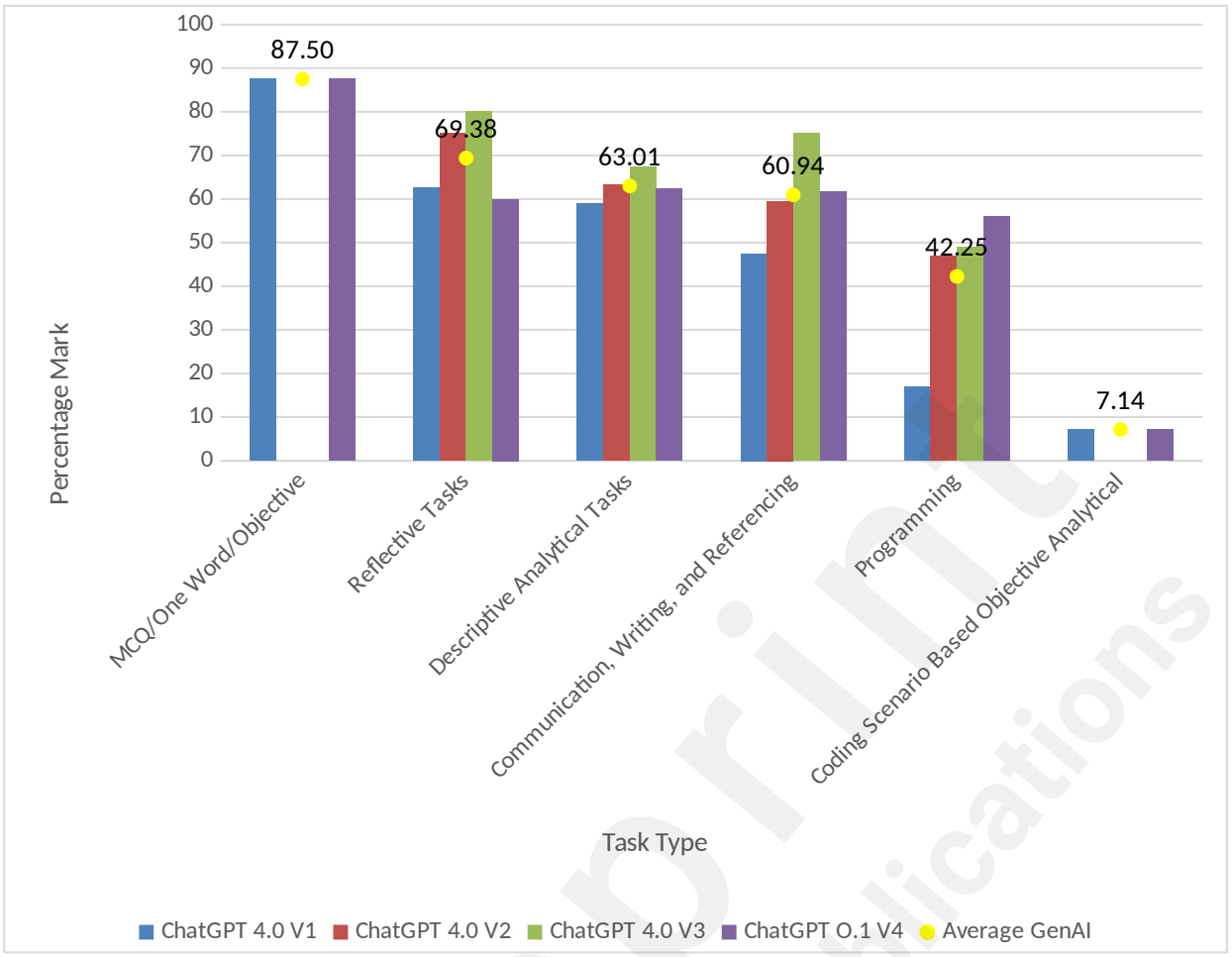


Figure 3: Overall Performance Comparison Across Assessment Task Types/Rubric Criterion

Discussion

This research aimed to pilot the evaluation of ChatGPT's performance across diverse DH and HIM assessment tasks, with a focus on promoting academic integrity. This analysis not only revealed the specific contexts in which ChatGPT excelled or fell short but also suggested broader implications for designing assessments that foster critical thinking, ethical practice, and real-world applicability. By situating ChatGPT's performance within the study's original goals, we were able to gain clearer insights into the potential and the limitations of GenAI in academic and professional environments.

The digital health case study and solution design assessments highlighted ChatGPT's limitations in addressing complex, context-specific tasks. While AI submissions consistently excelled in identifying foundational frameworks and generating structured content, they lacked the depth and contextualisation needed to create impactful, tailored solutions for specific patient populations. ChatGPT's inability to establish strong links between identified gaps and proposed solutions reflects broader challenges in maintaining depth and context. This finding aligns with observations from prior studies in other disciplines, which also reported similar shortcomings in AI-generated outputs, despite their polished language, coherent structure, and high grammatical accuracy [16,34]. Additionally, the absence of visual aids, such as patient journey maps or flowcharts, further limited the effectiveness of AI-generated responses in conveying complex ideas, thereby highlighting generative AI's limitations in producing detailed and meaningful illustrations [35]. The findings, therefore, underscore the critical role of human expertise in refining and contextualising proposals. When designing assessments for DIGHIM, tasks should require students to engage in deep

contextualisation, patient-specific scenarios and dynamic problem solving. This includes requiring students to integrate real-world data, include graphical visualisations, and justify decisions through detailed case-specific analyses, which can better evaluate their ability to apply theoretical knowledge in practice. Such approaches can ensure that assessments are not only aligned with professional standards but also resist over-reliance on AI tools, preserving the integrity of the learning process and fostering the development of essential critical thinking skills.

The reflective assessment demonstrated ChatGPT's strongest performance across all tasks, with later versions, particularly 4.0 V3, achieving the highest score (80%) and even surpassing marked student submissions in structure, coherence, and adherence to reflective frameworks. This echoes findings from dental education assessments, where ChatGPT demonstrated high performance in creating structured and coherent portfolio submissions [17]. This raises important questions about the role of generative AI in assessments designed to evaluate critical thinking and personal insight. While ChatGPT excelled in producing well-structured reports with clear language and systematic use of evidence, it sometimes lacked genuine depth, contextualisation, and personalised reflection - elements that are core to developing HIM professional competencies. The results, therefore, highlight the need to rethink the design of reflective assessments to maintain their authenticity and ensure they effectively evaluate students' skills. Reflective assessments should prioritise tasks that require personalised insights, unique contextual connections, and critical evaluation of experiences. Incorporating dynamic, scenario-based reflections, peer interactions, or requiring students to integrate real-time, context-specific observations can reduce reliance on AI-generated responses and emphasise genuine engagement [36,37]. This approach not only preserves academic integrity but also ensures that assessments continue to foster the critical, reflective skills essential for professional development.

Objective questions, such as multiple-choice and true/false tasks, showcased ChatGPT's strongest performance across all assessments, with consistently high accuracy rates and minimal errors. This aligns with the strengths of AI models in processing structured, rule-based tasks that require factual recall and logical reasoning. Similar findings have been reported in multiple evaluations of the performance of generative AI tools on objective or short answer assessments, reinforcing their reliability in domains where responses are clearly defined and less reliant on contextual interpretation [38–41]. The strong performance of AI in objective tasks raises concerns about the reliability of such assessments in evaluating student learning. If AI can reliably and accurately solve objective or knowledge recall questions, these tasks may no longer serve as a robust measure of individual student knowledge or skills, especially in an online environment. To ensure academic integrity and meaningful assessment outcomes, educators should consider integrating higher-order thinking components into objective assessments. For example, objective tasks could be complemented with reflective or explanatory questions that require students to justify their answers, provide reasoning, or discuss their thought processes. These additions would challenge students to engage more deeply with the material while mitigating the risk of AI dominating these tasks entirely [42]. Incorporating real-world context and multi-step problem-solving into objective questions can make them more authentic and aligned with learning objectives.

The clinical coding scenario-based questions exemplified an assessment activity, where a clear performance gap between AI and student submissions was found, particularly in tasks requiring contextual understanding and detailed application of clinical coding rules. While AI performed comparably to students in objective questions, it lacked the nuanced comprehension needed for coding scenarios, reflected in the significant difference in scores compared to the top-performing student. This aligns with findings from the USA, where ChatGPT 4 has been found to be effective in coding simple, single diagnoses using frequently extracted codes [43,44]. Soroush et el. [44], who

used more complex patient data derived from electronic medical records to test the coding efficacy of LLMs, found that while GPT-4 showed the highest exact match, it was only 33.9% for the ICD-10-CM, the USA's Clinical Modification of the International Classification of Diseases (ICD). Soroush et al.[44] (page 8) concluded that LLMs, including ChatGPT 4, do "not have a complete internal representation of medical coding rules" and are currently inappropriate for medical coding. This highlights the challenge of applying health classification logic beyond basic lookup tasks and underscores the need for assessments that emphasise contextual reasoning and real-world coding applications. While AI may support simpler, objective components of health classification assessments, its limitations reinforce the necessity for students to develop independent clinical coding proficiency and critical evaluation skills.

The findings from the SQL programming tasks revealed limitations in ChatGPT's ability to execute technical assessments requiring precision and contextual understanding. While later versions of ChatGPT showed incremental improvements, challenges persisted across all versions. Issues such as misaligned table and column usage, reliance on input instructions, and a lack of domain-specific knowledge of health data conventions frequently led to incomplete or inaccurate outputs. For example, ChatGPT struggled to handle demographic breakdowns, diagnosis descriptions, and complex multi-step queries, even when schema and rubric guidance were provided. Additionally, persistent errors like misspellings and the absence of real-time testing and debugging capabilities compounded its inability to produce reliable results. These findings contrast with prior research demonstrating high performance in SQL code generation by large language models [45]. In our study, however, AI-generated SQL responses performed significantly worse, suggesting that SQL coding in health databases requires a greater understanding of domain-specific knowledge, structured query logic, and data conventions. Past research has highlighted the critical role of contextual awareness and iterative testing in refining SQL code generation for generative AI models [46]. For educators, this underscores the importance of designing programming assessments that test problem-solving skills beyond basic syntax or query construction. Incorporating tasks that demand real-world application, debugging, and iterative testing within live environments would ensure that students develop practical, job-ready competencies that cannot be fully replicated by AI. Furthermore, assessments should include elements that require deeper interpretation and contextualisation of health data, encouraging students to demonstrate both technical accuracy and analytical reasoning.

ChatGPT's performance in communication, writing, and referencing tasks demonstrated notable strengths in structure, language clarity, and adherence to academic conventions, particularly in later versions like V3. However, issues with word length, depth, contextual relevance, and repetition persisted, especially in earlier versions, aligning with findings from a generative AI performance evaluation study for business education [16]. They found that for longer academic writing tasks requiring 1500 words, AI-generated responses consistently fell short, with scores ranging between 40% and 77% due to a lack of depth, context, and critical engagement. While AI-generated submissions exhibited strong grammatical quality and high Grammarly scores, they frequently lacked academic substance, critical analysis, and nuanced argumentation. Chaudhry et al. [16] also observed the presence of hallucinated references and inconsistencies in citation formatting, which undermined the credibility of submissions, a challenge extensively documented in previous research [47,48]. These findings suggest that while ChatGPT can support foundational academic writing, it may struggle with nuanced tasks requiring original thought, contextual understanding, and argumentation. To maintain academic integrity, educators should design assessments that require critical engagement with sources, unique application of concepts, and personalised insights, elements that AI cannot easily replicate. Integrating tasks that demand deeper analysis, precise referencing, and clear connections between evidence and arguments will ensure that students demonstrate authentic writing and reasoning skills that AI tools alone cannot fulfil.

‘Chain prompting’, where prompts were refined and built upon in stages, played a critical role in improving ChatGPT’s performance, particularly in descriptive and reflective tasks, with later versions like V3 benefiting from rubric-driven guidance. Similar improvements were observed in a pharmacy structured assessment, where chain prompting significantly boosted performance in knowledge recall-based tasks [41]. This highlights the effectiveness of structured, multi-step prompting in optimising AI outputs across different assessment formats, particularly those requiring factual accuracy and well-defined responses.

Overall, ChatGPT version o1 preview (v4 o1) underperformed in descriptive tasks compared to V3, although its performance matched or improved in simpler, objective tasks such as multiple-choice questions. This indicates that higher AI versions do not necessarily guarantee better outcomes, particularly for complex, context-driven assessments. While previous studies have shown that GPT-4 outperforms GPT-3.5 in medical licensing examinations [49], there remains a lack of comparative research evaluating the impact of newer o1 models on academic assessment performance. Although earlier findings reported strong performance in basic sciences [50], this was not consistently reflected in our study, where ChatGPT o1 demonstrated no clear improvement and, in some cases, a decline in descriptive tasks. Students with stronger prompting skills may benefit in tasks aligned with AI strengths, which could potentially widen academic disparities. However, the findings provide no consistent evidence that newer versions like ChatGPT o1 enhance performance in tasks requiring critical thinking and reflection. These results highlight the importance for educators to design assessments that emphasise originality, contextual understanding, and higher-order thinking, rather than tasks that can be easily completed with AI assistance.

Significance and implications

The findings from this pilot study highlight the importance of designing assessments that align with the evolving roles in Health Information Management and Digital Health while fostering the ethical and practical use of generative AI. Health Information Managers play a crucial role in health data management, analytics, and health ICT [28], as Digital Health increasingly integrates disciplines such as clinical care, data science, and information technology[51], it is essential for educational assessments to reinforce interdisciplinary skills that align with the evolving demands of these domains [52,53]. Recent evidence highlights that traditional assessment methods in higher education are increasingly ineffective in GenAI-facilitated learning environments, necessitating redesigned assessments that promote critical thinking, creativity, and lifelong learning skills [25].

Assessments should emphasise tasks that require contextual understanding, technical proficiency, and decision-making. Furthermore, these assessments should integrate AI as a tool to augment, rather than a substitute for critical thinking. For example, students could refine AI-generated clinical coding or SQL outputs to align with standards-based frameworks, critically evaluate AI-generated business proposals for depth and contextual relevance or integrate visualisations like patient journey maps into health information reports. These approaches not only build essential competencies such as critical analysis, programming, and communication but also encourage students to navigate AI's limitations and responsibly incorporate its outputs into their work, promoting the ethical use of generative AI to support, rather than replace, student efforts.

In DIGHIM education, AI use in assessments should be carefully structured to align with professional competencies while maintaining academic integrity. Some tasks should avoid reliance on AI, particularly those requiring personal reflections, ethical decision-making, and critical thinking

in complex digital health scenarios. Other tasks can adopt AI for efficiency, such as assisting in summarising health policies, generating structured reports, or organising data-driven insights. Finally, AI-generated content can be adapted in assessments that require students to critically refine, validate, and contextualise AI outputs, such as evaluating AI-generated clinical coding recommendations, refining SQL queries for health databases, or assessing the applicability of AI-generated digital health solutions within industry frameworks. Additionally, educators and students must ensure privacy by refraining from supplying generative AI tools with copyrighted or sensitive health data and should receive structured training on the responsible and ethical use of AI in digital health and HIM to support professional readiness.

These approaches prepare students for real-world applications of generative AI in HIM and Digital Health roles, where such tools may be used to assist with technical tasks like data analysis, report generation, and strategic planning. By designing assessments that both test and build skills in these areas while fostering responsible AI use, educators can equip graduates to navigate the interdisciplinary and technology-driven landscape of modern digital healthcare effectively.

Table 8 provides a summary of important recommendations.

Table 8: Recommendations for ethical assessment design in DIGHIM in the AI era

No.	Key findings	Recommendations
1	AI struggled with complex, context-specific tasks	Design assessments that require deep contextualisation, dynamic problem-solving, and patient-specific scenarios to reduce over-reliance on AI.
2	Reflective assessments were AI's strongest area, even outperforming students in structure and coherence	Improve reflective assessments by requiring personalised insights, scenario-based reflections, peer interactions, and integration of real-time experiences.
3	AI performed well on objective questions but may have compromised originality and contextual understanding	Enhance objective assessments with higher-order thinking components, such as explanatory questions, justification of answers, and real-world problem-solving to ensure genuine learning.
4	Clinical coding scenario tasks revealed significant AI limitations	Develop clinical coding assessments that emphasise reasoning, logic, and real-world applications rather than simple lookup tasks, ensuring AI cannot replace student competency.
5	SQL programming tasks for health databases required real-time testing, debugging, and domain expertise	Design hands-on programming assessments that require students to iteratively test and debug queries in live database environments. Emphasise contextual understanding, scenario-based problem-solving, and adherence to health database conventions to ensure students develop practical technical skills essential for real-

		world applications.
6	AI performed well in structured writing, but lacked depth, critical argumentation, and contextual understanding	Ensure writing tasks focus on original thought, deeper argumentation, and nuanced critical engagement. Require students to provide their own contextual reasoning rather than relying solely on AI-generated text. Adopt an AI-integrated approach where students critically evaluate, refine, and justify AI-generated outputs. Require them to validate AI outputs against academic sources.
7	Chain prompting improved performance in structured tasks, but not in complex reasoning tasks	Utilize structured, rubric-driven prompts to improve AI-generated responses, while ensuring assessments test deeper analysis and application of knowledge rather than surface-level synthesis.
8	Higher AI versions did not always improve performance in descriptive tasks	Monitor AI version capabilities, as newer versions do not necessarily enhance performance in descriptive tasks. Educators should test different AI models before integrating them into learning workflows.
9	AI detection methods were unreliable and may have produced false positives	Avoid over-reliance on AI detection tools; instead, design assessments that emphasize originality, reasoning, and personalised insights to differentiate human work from AI-generated responses.
10	Generative AI struggled with producing accurate citations and referencing	Strengthen academic writing assessments by requiring precise referencing techniques and verification of sources. Implement tasks where students critique AI-generated references for accuracy.
11	Students with strong AI prompting skills may have gained an academic advantage	Provide standardised AI literacy training to ensure equitable access to effective AI use while maintaining academic integrity and fairness.
12	Generative AI had limitations in illustration and visual content generation	Require students to create their own graphical illustrations (e.g., patient journey maps, workflow diagrams) rather than relying on AI-generated visuals, reinforcing deeper understanding.
13	Ethical concerns existed regarding AI's role in assessments	Develop institutional policies that regulate ethical AI use in assessments, ensuring AI enhances learning rather than undermining student skill development.
14	Detailed rubrics enabled AI to produce better responses and gain higher marks	Use simple, structured rubrics to guide learning without disclosing excessive detail that could allow AI to game the assessment. Balance transparency with the need to assess genuine understanding.
15	Maintaining data privacy when using	Train students to refrain from supplying

	generative AI tools was essential	copyrighted or sensitive health data to generative AI tools, ensuring compliance with data protection policies.
16	AI-generated responses varied in quality depending on the assessment type, performing well in knowledge retention or procedural tasks but poorly in assessments requiring ethical reasoning, deeper personal reflection, or nuanced judgment.	Implement AI-integrated assessment strategies that distinguish between tasks that should avoid AI reliance (e.g., ethical decision-making, personal reflections), adopt AI for efficiency (e.g., summarizing health policies), and adapt AI outputs for deeper engagement (e.g., refining and critically analysing AI-generated solutions).

Limitations and future work

This study had several limitations. First, as a pilot investigation, the findings should be interpreted as preliminary, offering directional insights into AI performance across DIGHIM assessment types rather than definitive conclusions. Second, the quasi-experimental design lacks the random assignment of traditional experimental designs, potentially limiting the generalisability of findings. Third, assessments were evaluated using specific markers, whose individual grading preferences may have introduced subtle biases despite the blinded review process and use of rubrics. Fourth, while the study examined various assessment types, the limited number of assessment samples may have affected the robustness of conclusions for these specific tasks. Additionally, the inability to replicate real-time AI usage scenarios by students in dynamic educational settings could understate the challenges or benefits of generative AI in practice. Last, we cannot determine with certainty whether students used AI tools in their original submissions, which may influence comparisons between student and AI-generated responses.

An additional limitation is that the study was not designed to systematically examine the detectability of AI-generated content. While markers occasionally inferred AI authorship based on features such as formulaic phrasing, generic recommendations, or inconsistent contextual alignment, these impressions were anecdotal and at times inaccurate, with some student submissions misidentified as AI-generated. Such false positives are well documented in the literature [16,27,54] and highlight the unreliability of human detection methods. Future research should explicitly investigate the accuracy of markers in distinguishing AI- from student-generated work and consider how assessment design can emphasise originality, critical thinking, and personalised insights to reduce reliance on detection and strengthen academic integrity.

Future research should also focus on exploring AI's performance across diverse assessment formats. Investigating longitudinal impacts of AI tools on student learning outcomes, skill development, and academic integrity could provide deeper insights. Moreover, research should investigate AI's role in dynamic scenarios, such as patient data analysis, health system design, or real-world interoperability challenges, to better align assessments with industry demands. Lastly, privacy and ethical considerations, such as ensuring AI tools are not trained on sensitive or copyrighted content, warrant further exploration to guide responsible AI implementation in higher education.

.

Conclusions

This pilot study has highlighted the potential and limitations of ChatGPT in Health Information Management and Digital Health assessments. While generative AI demonstrates strengths in structured tasks and foundational content generation, it struggles with contextualisation, depth, and the technical precision required for assessments requiring independent critical thinking. These

findings emphasise the need for carefully designed assessments that integrate AI ethically and prioritise tasks that require human judgment and contextual understanding to ensure meaningful learning outcomes and academic integrity.

Acknowledgements

The authors sincerely thank the staff of the Digital Health and Information Management cluster, the participating markers, and the academic leadership of the School of Psychology and Public Health at La Trobe University for their support of this study.

Funding

The research was funded through La Trobe University's internal AI in teaching grant scheme.

Author Contribution Statement (CReDIT)

T.A. Wani: Conceptualisation, Data curation, Formal analysis, Investigation, Methodology, Validation, Project administration, Resources, Visualisation, Validation, Supervision, Writing – original draft
M. Liem: Conceptualisation, Data curation, Methodology, Investigation, Project administration, Validation, Resources, Writing – review and editing
N. Prasad: Conceptualisation, Data curation, Methodology, Investigation, Project administration, Resources, Writing – review and editing
K. Robinson: Formal Analysis, Investigation, Supervision, Writing – review & editing.
A. Nexhip: Data curation, Formal analysis, Investigation, Writing – review & editing.
M. Tassos: Investigation, Formal analysis, Validation, Writing – review & editing.
S. Gjorgioski: Investigation, Formal analysis, Validation, Writing – review & editing.
U.R. Khan: Formal analysis, Validation, Writing – review & editing.
J. Boyd: Formal analysis, Validation, Writing – review & editing.
M. Riley: Conceptualisation, Data Curation, Formal Analysis, Validation, Methodology, Project administration, Supervision, Resources, Writing – review and editing

Conflicts of Interest

None declared.

Abbreviations

References

1. Makridakis S. The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms. Futures 2017 June 1;90:46–60. doi: 10.1016/j.futures.2017.03.006

2. ISO/IEC 22989:2022(en), Information technology — Artificial intelligence — Artificial intelligence concepts and terminology. Available from: https://www.iso.org/obp/ui/#iso:std:iso-iec:22989:ed-1:v1:en [accessed May 16, 2025]

3. Marr B. The 4 Types Of Generative AI Transforming Our World. Forbes. 2024. Available from: https://www.forbes.com/sites/bernardmarr/2024/04/29/the-4-types-of-generative-ai-transforming-our-world/ [accessed May 16, 2025]

4. Hasanein AM, Sobaih AEE. Drivers and Consequences of ChatGPT Use in Higher Education: Key Stakeholder Perspectives. Eur J Investig Health Psychol Educ 2023 Nov 9;13(11):2599–2614. PMID:37998071

5. Kasneci E, Sessler K, Küchemann S, Bannert M, Dementieva D, Fischer F, Gasser U, Groh G, Günnemann S, Hüllermeier E, Krusche S, Kutyniok G, Michaeli T, Nerdel C, Pfeffer J, Poquet O, Sailer M, Schmidt A, Seidel T, Stadler M, Weller J, Kuhn J, Kasneci G. ChatGPT for good? On opportunities and challenges of large language models for education. Learn Individ Differ 2023 Apr 1;103:102274. doi: 10.1016/j.lindif.2023.102274

6. Nazi ZA, Peng W. Large Language Models in Healthcare and Medical Domain: A Review. Informatics Multidisciplinary Digital Publishing Institute; 2024 Sept;11(3):57. doi: 10.3390/informatics11030057

7. Gao CA, Howard FM, Markov NS, Dyer EC, Ramesh S, Luo Y, Pearson AT. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. Npj Digit Med Nature Publishing Group; 2023 Apr 26;6(1):1–5. doi: 10.1038/s41746-023-00819-6

8. Lund BD, Wang T, Mannuru NR, Nie B, Shimray S, Wang Z. ChatGPT and a new academic reality: Artificial Intelligence-written research papers and the ethics of the large language models in scholarly publishing. J Assoc Inf Sci Technol 2023;74(5):570–581. doi: 10.1002/asi.24750

9. Lin Z. Why and how to embrace AI such as ChatGPT in your academic life. R Soc Open Sci Royal Society; 2023 Aug 23;10(8):230658. doi: 10.1098/rsos.230658

10. Deng J, Lin Y. The Benefits and Challenges of ChatGPT: An Overview. Front Comput Intell Syst 2022;2(2):81–83. doi: 10.54097/fcis.v2i2.4465

11. Pradana M, Elisa HP, Syarifuddin S. Discussing ChatGPT in education: A literature review and bibliometric analysis. Cogent Educ Cogent OA; 2023 Dec 11;10(2):2243134. doi: 10.1080/2331186X.2023.2243134

12. Baig MI, Yadegaridehkordi E. ChatGPT in the higher education: A systematic literature review and research challenges. Int J Educ Res 2024 Jan 1;127:102411. doi: 10.1016/j.ijer.2024.102411

13. Rasul T, Nair S, Kalendra D, Robin M, Santini F de O, Ladeira WJ, Sun M, Day I, Rather RA, Heathcote L. The role of ChatGPT in higher education: Benefits, challenges, and future research directions. J Appl Learn Teach 2023 May 10;6(1):41–56. doi: 10.37074/jalt.2023.6.1.29

14. Rawas S. ChatGPT: Empowering lifelong learning in the digital age of higher education. Educ Inf Technol 2024 Apr 1;29(6):6895–6908. doi: 10.1007/s10639-023-12114-8

15. Nikolic S, Daniel ,Scott, Haque ,Rezwanul, Belkina ,Marina, Hassan ,Ghulam M., Grundy ,Sarah, Lyden ,Sarah, Neal ,Peter, and Sandison C. ChatGPT versus engineering

education assessment: a multidisciplinary and multi-institutional benchmarking and analysis of this generative artificial intelligence tool to investigate assessment integrity. Eur J Eng Educ Taylor & Francis; 2023 July 4;48(4):559–614. doi: 10.1080/03043797.2023.2213169

16. Chaudhry IS, Sarwary SAM, El Refae GA, Chabchoub H. Time to Revisit Existing Student's Performance Evaluation Approach in Higher Education Sector in a New Era of ChatGPT — A Case Study. Cogent Educ Cogent OA; 2023 Dec 31;10(1):2210461. doi: 10.1080/2331186X.2023.2210461

17. Ali K, Barhom N, Tamimi F, Duggal M. ChatGPT—A double-edged sword for healthcare education? Implications for assessments of dental students. Eur J Dent Educ 2024;28(1):206–211. doi: 10.1111/eje.12937

18. Hadi Mogavi R, Deng C, Juho Kim J, Zhou P, D. Kwon Y, Hosny Saleh Metwally A, Tlili A, Bassanelli S, Bucchiarone A, Gujar S, Nacke LE, Hui P. ChatGPT in education: A blessing or a curse? A qualitative study exploring early adopters' utilization and perceptions. Comput Hum Behav Artif Hum 2024 Jan 1;2(1):100027. doi: 10.1016/j.chbah.2023.100027

19. Sallam M, Salim N, Barakat M, Al-Tammemi A. ChatGPT applications in medical, dental, pharmacy, and public health education: A descriptive study highlighting the advantages and limitations. Narra J 2023 May 1;3(1):e103–e103. doi: 10.52225/narra.v3i1.103

20. Akiba D, Fraboni MC. AI-Supported Academic Advising: Exploring ChatGPT's Current State and Future Potential toward Student Empowerment. Educ Sci Multidisciplinary Digital Publishing Institute; 2023 Sept;13(9):885. doi: 10.3390/educsci13090885

21. Farhud DD, Zokaei S. Ethical Issues of Artificial Intelligence in Medicine and Healthcare. Iran J Public Health 2021 Oct 27; doi: 10.18502/ijph.v50i11.7600

22. Rahman MM, Watanobe Y. ChatGPT for Education and Research: Opportunities, Threats, and Strategies. Appl Sci Multidisciplinary Digital Publishing Institute; 2023 Jan;13(9):5783. doi: 10.3390/app13095783

23. Neumann M, Rauschenberger M, Schön E-M. "We Need To Talk About ChatGPT": The Future of AI and Higher Education. IEEE; 2023. doi: 10.25968/opus-2978

24. AlAfnan MA, Dishari S, Jovic M, Lomidze K. ChatGPT as an Educational Tool: Opportunities, Challenges, and Recommendations for Communication, Business Writing, and Composition Courses. J Artif Intell Technol 2023 Mar 6;3(2):60–68. doi: 10.37965/jait.2023.0184

25. Weng X, Xia Q, Gu M, Rajaram K, Chiu TKF. Assessment and learning outcomes for generative AI in higher education: A scoping review on current research status and trends. Australas J Educ Technol 2024 Oct 18;40(6):37–55. doi: 10.14742/ajet.9540

26. Foltynek T, Bjelobaba S, Glendinning I, Khan ZR, Santos R, Pavletic P, Kravjar J. ENAI Recommendations on the ethical use of Artificial Intelligence in Education. Int J Educ Integr BioMed Central; 2023 Dec;19(1):1–4. doi: 10.1007/s40979-023-00133-4

27. Elkhatat AM, Elsaid K, Almeer S. Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. Int J Educ Integr BioMed Central; 2023 Dec;19(1):1–16. doi: 10.1007/s40979-023-00140-5

28. Gjorgioski S, Riley M, Lee J, Prasad N, Tassos M, Nexhip A, Richardson S, Robinson K. Workforce survey of Australian health information management graduates, 2017–2021: A 5-year follow-on study. Health Inf Manag J SAGE Publications Ltd STM; 2023;54(1):43–54. doi: 10.1177/18333583231197936

29. Riley M, Robinson K, Prasad N, Gleeson B, Barker E, Wollersheim D, Price J. Workforce survey of Australian graduate health information managers: Employability, employment, and knowledge and skills used in the workplace. Health Inf Manag J SAGE Publications Ltd STM; 2020 May 1;49(2–3):88–98. doi: 10.1177/1833358319839296

30. Temsah M-H, Aljamaan F, Malki KH, Alhasan K, Altamimi I, Aljarbou R, Bazuhair F, Alsubaihin A, Abdulmajeed N, Alshahrani FS, Temsah R, Alshahrani T, Al-Eyadhy L, Alkhateeb SM, Saddik B, Halwani R, Jamal A, Al-Tawfiq JA, Al-Eyadhy A. ChatGPT and the Future of Digital Health: A Study on Healthcare Workers' Perceptions and Expectations. Healthc Basel Switz 2023 June 21;11(13):1812. PMID:37444647

31. Tudor Car L, Kyaw BM, Nannan Panday RS, van der Kleij R, Chavannes N, Majeed A, Car J. Digital Health Training Programs for Medical Students: Scoping Review. JMIR Med Educ 2021 July 21;7(3):e28275. PMID:34287206

32. Eysenbach G. The Role of ChatGPT, Generative Language Models, and Artificial Intelligence in Medical Education: A Conversation With ChatGPT and a Call for Papers. JMIR Med Educ 2023 Mar 6;9(1):e46885. doi: 10.2196/46885

33. Meo SA, Al-Masri AA, Alotaibi M, Meo MZS, Meo MOS. ChatGPT Knowledge Evaluation in Basic and Clinical Medical Sciences: Multiple Choice Question Examination-Based Performance. Healthc Basel Switz 2023 July 17;11(14):2046. PMID:37510487

34. Kolade O, Owoseni A, Egbetokun A. Is AI changing learning and assessment as we know it? Evidence from a ChatGPT experiment and a conceptual framework. Heliyon 2024 Feb 29;10(4):e25953. doi: 10.1016/j.heliyon.2024.e25953

35. Moin KA, Nasir AA, Petroff DJ, Loveless BA, Moshirfar OA, Hoopes PC, Moshirfar M, Moin KA, Nasir AA, Petroff DJ, Loveless BA, Moshirfar O, Sr PH, Moshirfar M. Assessment of Generative Artificial Intelligence (AI) Models in Creating Medical Illustrations for Various Corneal Transplant Procedures. Cureus Cureus; 2024 Aug 26;16. doi: 10.7759/cureus.67833

36. Powell S, Forsyth R. Generative AI and the implications for authentic assessment. Using Gener AI Eff High Educ Routledge; 2024. ISBN:978-1-003-48291-8

37. Salinas-Navarro DE, Vilalta-Perdomo E, Michel-Villarreal R, Montesinos L. Designing experiential learning activities with generative artificial intelligence tools for authentic assessment. Interact Technol Smart Educ Emerald Publishing Limited; 2024 May 6;21(4):708–734. doi: 10.1108/ITSE-12-2023-0236

38. Hersh W, Fultz Hollis K. Results and implications for generative AI in a large introductory biomedical and health informatics course. Npj Digit Med Nature Publishing Group; 2024 Sept 13;7(1):1–7. doi: 10.1038/s41746-024-01251-0

39. Morjaria L, Burns L, Bracken K, Ngo QN, Lee M, Levinson AJ, Smith J, Thompson P, Sibbald M. Examining the Threat of ChatGPT to the Validity of Short Answer Assessments in an

Undergraduate Medical Program. J Med Educ Curric Dev SAGE Publications Ltd STM; 2023 Jan 1;10:23821205231204178. doi: 10.1177/23821205231204178

40. Newton P, and Xiromeriti M. ChatGPT performance on multiple choice question examinations in higher education. A pragmatic scoping review. Assess Eval High Educ SRHE Website; 2024 Aug 17;49(6):781–798. doi: 10.1080/02602938.2023.2299059

41. Yang H, Hu M, Most A, Hawkins WA, Murray B, Smith SE, Li S, Sikora A. Evaluating accuracy and reproducibility of large language model performance on critical care assessments in pharmacy education. Front Artif Intell Frontiers; 2025 Jan 9;7:1514896. doi: 10.3389/frai.2024.1514896

42. Thanh BN, Vo DTH, Nhat MN, Pham TTT, Trung HT, Xuan SH. Race with the machines: Assessing the capability of generative AI in solving authentic assessments. Australas J Educ Technol 2023 Dec 22;39(5):59–81. doi: 10.14742/ajet.8902

43. Abdelgadir Y, Thongprayoon C, Miao J, Suppadungsuk S, Pham JH, Mao MA, Craici IM, Cheungpasitporn W. AI integration in nephrology: evaluating ChatGPT for accurate ICD-10 documentation and coding. Front Artif Intell 2024;7:1457586. PMID:39286549

44. Soroush A, Glicksberg BS, Zimlichman E, Barash Y, Freeman R, Charney AW, Nadkarni GN, Klang E. Large language models are poor medical coders—benchmarking of medical code querying. NEJM AI Massachusetts Medical Society; 2024;1(5):AIdbp2300040.

45. Pornphol P, Chittayasothorn S. Verification of Relational Database Languages Codes Generated by ChatGPT. Proc 2023 4th Asia Serv Sci Softw Eng Conf New York, NY, USA: Association for Computing Machinery; 2024. p. 17–22. doi: 10.1145/3634814.3634817

46. Carr N, Shawon FR, Jamil HM. An Experiment on Leveraging ChatGPT for Online Teaching and Assessment of Database Students. 2023 IEEE Int Conf Teach Assess Learn Eng TALE 2023. p. 1–8. doi: 10.1109/TALE56641.2023.10398239

47. Walters WH, Wilder EI. Fabrication and errors in the bibliographic citations generated by ChatGPT. Sci Rep Nature Publishing Group; 2023 Sept 7;13(1):14045. doi: 10.1038/s41598-023-41032-5

48. Williams A. Comparison of generative AI performance on undergraduate and postgraduate written assessments in the biomedical sciences. Int J Educ Technol High Educ 2024 Sept 13;21(1):52. doi: 10.1186/s41239-024-00485-y

49. Liu M, Okuhara T, Chang X, Shirabe R, Nishiie Y, Okada H, Kiuchi T. Performance of ChatGPT Across Different Versions in Medical Licensing Examinations Worldwide: Systematic Review and Meta-Analysis. J Med Internet Res 2024 July 25;26(1):e60807. doi: 10.2196/60807

50. Jones N. ‘In awe’: scientists impressed by latest ChatGPT model o1. Nature 2024 Oct 1;634(8033):275–276. doi: 10.1038/d41586-024-03169-9

51. Meehan R. Health Informatics Workforce in the Digital Health Ecosystem. MEDINFO 2023 — Future Access IOS Press; 2024. p. 1226–1230. doi: 10.3233/SHTI231160

52. Butler-Henderson K, Gray K, Arabi S. Roles and Responsibilities of the Global Specialist

Digital Health Workforce: Analysis of Global Census Data. JMIR Med Educ 2024 July 25;10(1):e54137. doi: 10.2196/54137

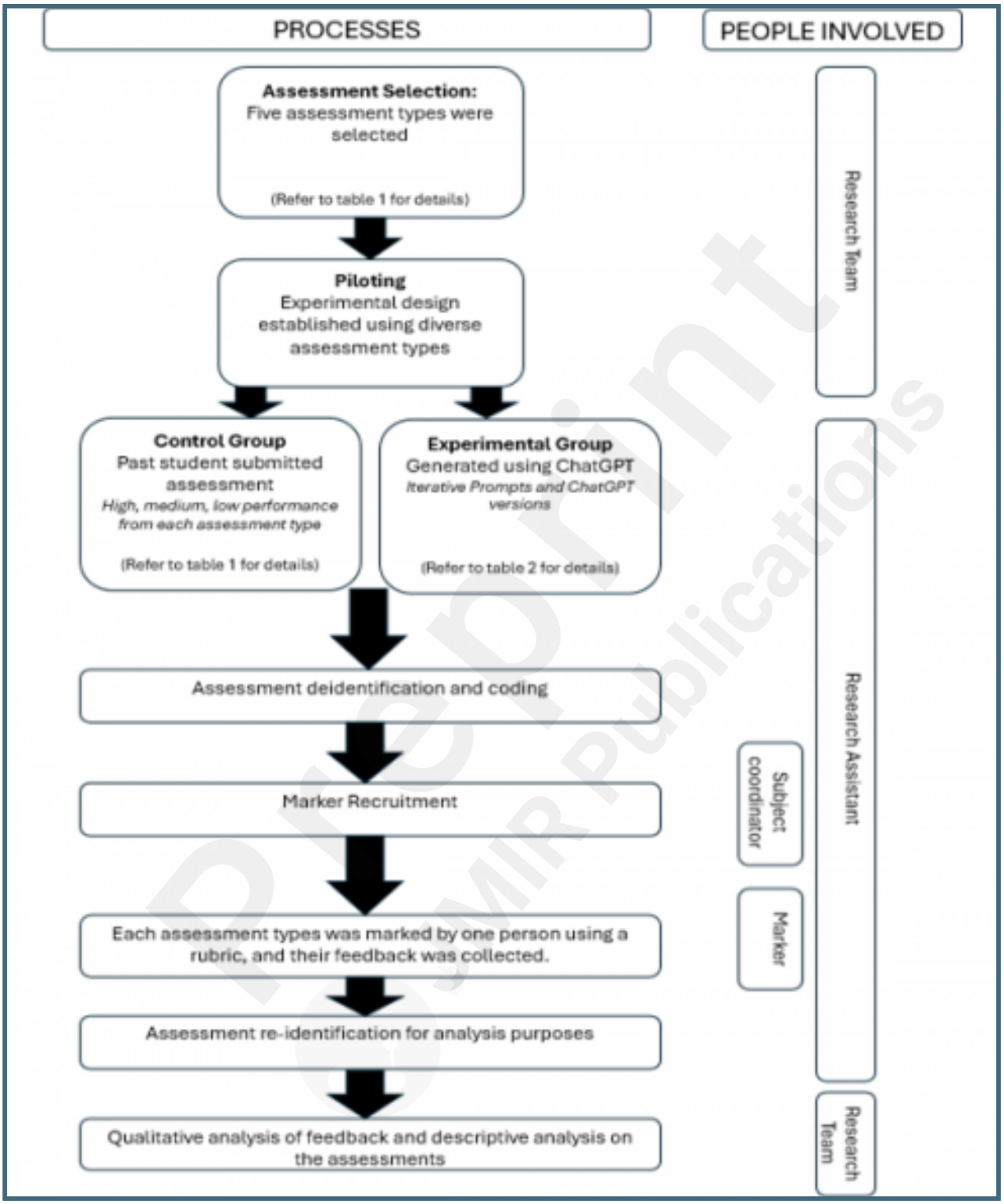
53. Keep M, Janssen A, McGregor D, Brunner M, Baysari MT, Quinn D, Shaw T. Mapping eHealth Education: Review of eHealth Content in Health and Medical Degrees at a Metropolitan Tertiary Institute in Australia. JMIR Med Educ 2021 Aug 19;7(3):e16440. PMID:34420920

54. Khalil M, Er E. Will ChatGPT Get You Caught? Rethinking of Plagiarism Detection. In: Zaphiris P, Ioannou A, editors. Learn Collab Technol Cham: Springer Nature Switzerland; 2023. p. 475–487. doi: 10.1007/978-3-031-34411-4_32

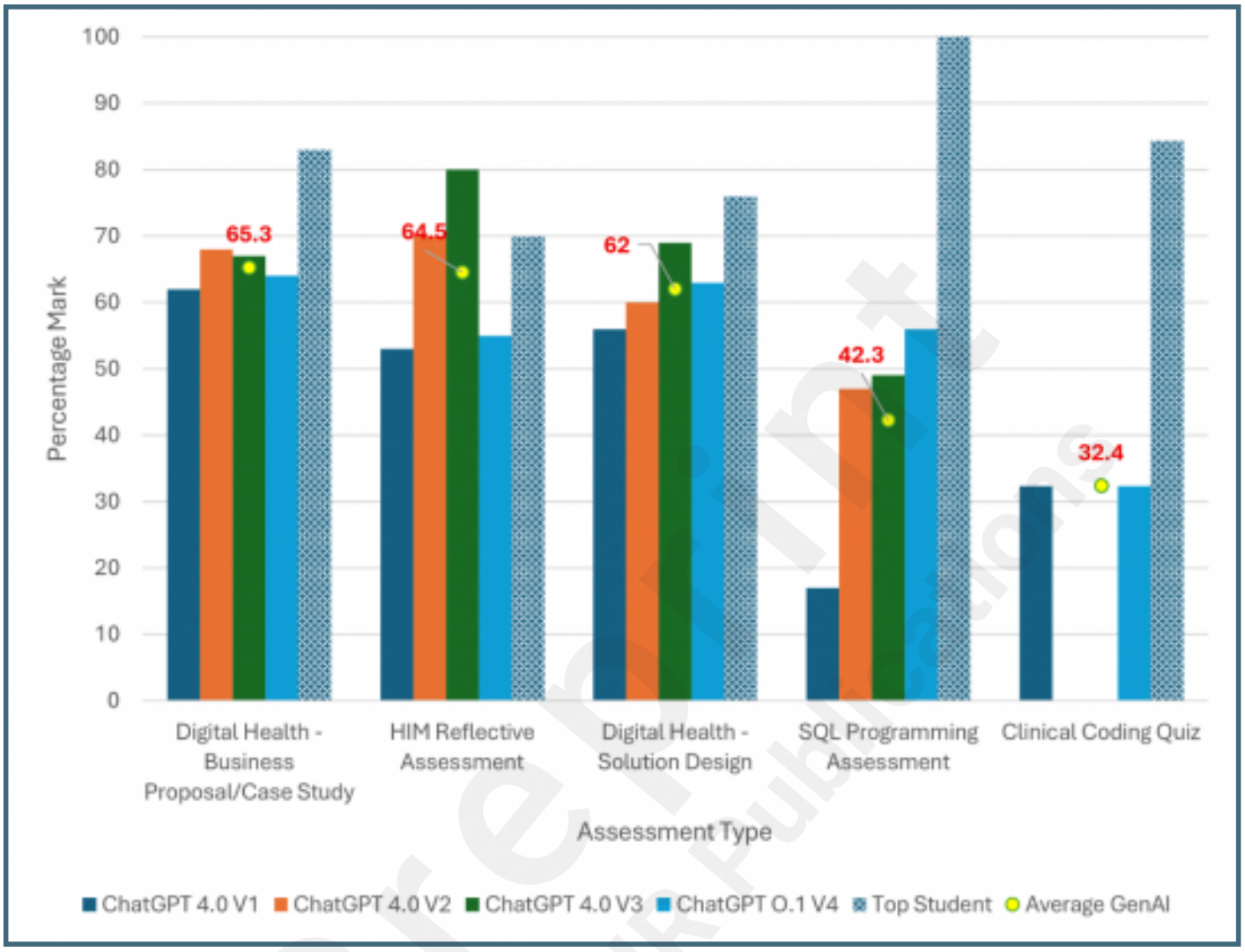
Supplementary Files

Figures

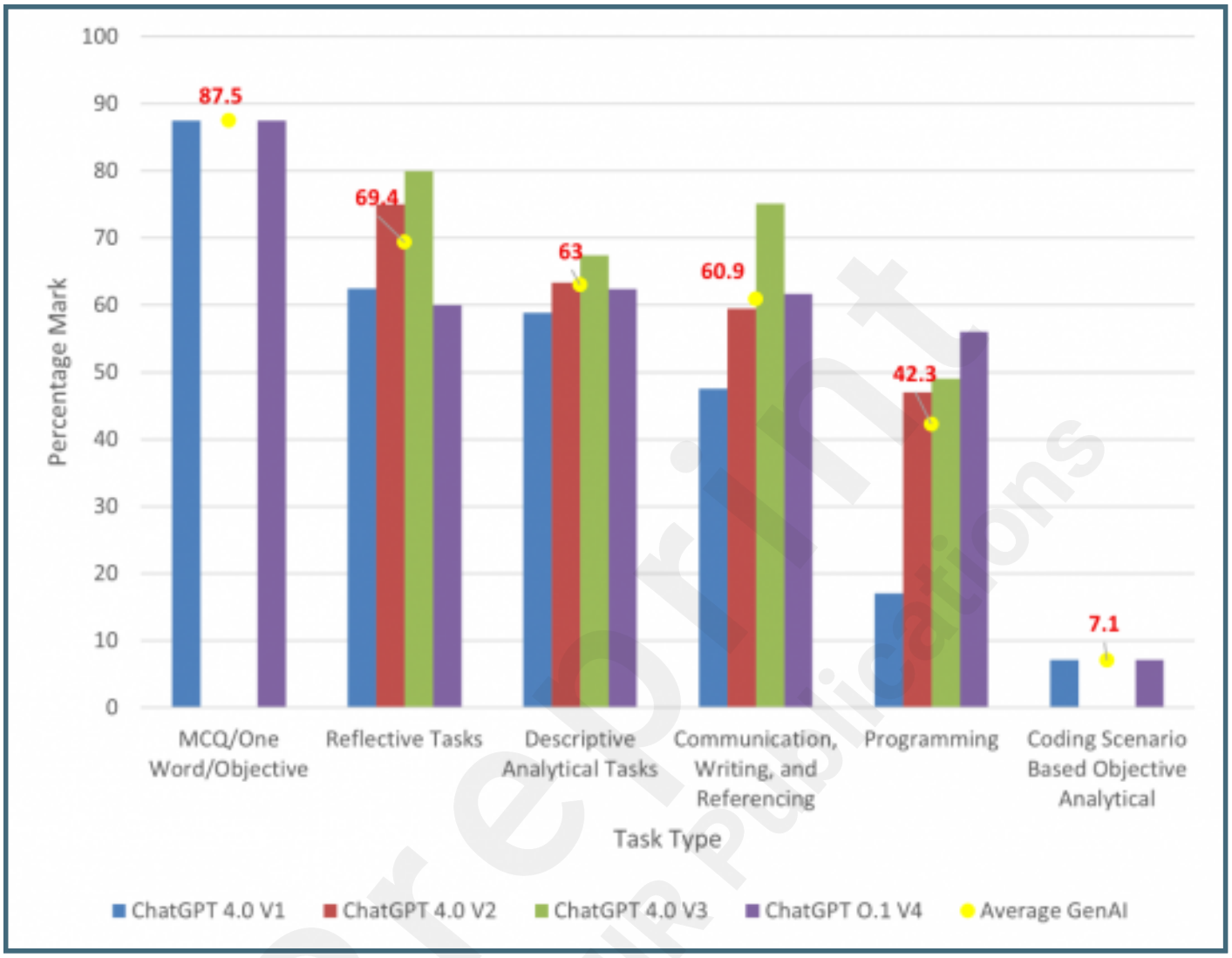
Overview of methodology.



Overall performance comparison across assessment types.



Overall performance comparison across assessment task types/rubric criterion.



Multimedia Appendixes

Detailed performance analysis by assessment type.
URL: http://asset.jmir.pub/assets/422aec6b0c417924cb83bbae50bb80e0.xlsx