Engineering Educators Bringing the World Together
2025 ASEE Annual Conference & Exposition
Palais des congrès de Montréal, Montréal, QC • June 22–25, 2025
ASEE
Paper ID #48701

Evaluating the Effectiveness of Generative AI for Automated Quiz Creation: A Case Study

Ms. Venkata Alekhya Kusam, University of Michigan - Dearborn

Venkata Alekhya Kusam received her Master's degree in Computer and Information Science from the University of Michigan-Dearborn. She is currently working as an AI developer and will be joining the University of British Columbia as a PhD student in Computer Science. Her research interests include Explainable AI (XAI), the integration of AI in education, and the development of human-centric AI systems to enhance learning and accessibility.

Zheng Song, University of Michigan - Dearborn

Dr. Song received his second PhD in Computer Science (with a focus on distributed systems and software engineering) from Virginia Tech USA in 2020, and the first PhD (with a focus on wireless networking and mobile computing) from Beijing University of Posts and Telecommunications China in 2015. He worked as a software engineer at Sina for one year after he graduated as a master from China Agriculture University in 2009. He received the Best Paper Award from IEEE Edge in 2019.

Khalid Kattan, University of Michigan - Dearborn

Dr. Khalid Kattan received his Ph.D. from Wayne State University in 2019 in Artificial Intelligence with a focus on Evolutionary Computation, Genetic and Cultural Algorithms. He has over 20 of industrial experience. He is passionate about AI and its impacts in learning.

Dr. Bruce R Maxim, University of Michigan - Dearborn

Bruce R. Maxim has worked as a software engineer, project manager, professor, author, and consultant for more than forty years. His research interests include software engineering, human computer interaction, game design, social media, artificial intellig

Evaluating the Effectiveness of Generative AI for Automated Quiz Creation: A Case Study

Venkata Alekhya Kusam, Zheng Song, Bruce Maxim, Khalid Kattan

{alekhyak, zhesong, bmaxim, kkattan}@umich.edu

Department of Computer and Information Science, University of Michigan-Dearborn, USA

Abstract

This paper presents an investigation into the use of Generative AI (GenAI), specifically ChatGPT, to automate quiz generation in higher education by conducting a case study in a graduate Artificial Intelligence (AI) course. The study aims to compare the quality and relevance of AI-generated quizzes with manually created ones, addressing a critical question in computer science education: Can Generative AI effectively support educators in creating assessments that align with course learning objectives?

We conducted the study in a graduate-level AI course, which involved 47 students, one instructor and one teaching assistant (TA). The manual quizzes were created by the instructor who is knowledgeable about the domain, while the GenAI-based quizzes were created by the TA who is knowledgeable about GenAI. Both ways of quiz generation focused on the same course topics, including machine learning, neural networks, and search algorithms. These quizzes were then assessed through a combination of qualitative and quantitative measures, focusing on alignment with learning outcomes, question relevance, and student performance. This approach allowed for a comprehensive comparison of the effectiveness of AI-generated quizzes versus those created manually.

Our results show that AI-generated quizzes excel in clarity, consistency, and efficiency but tend to be easier and less effective at differentiating student performance. In contrast, manually created quizzes offer greater depth, better alignment with course objectives, and foster critical thinking, though they require more effort to design. These findings offer evidence-based insights into the strengths and limitations of AI in educational assessment. To address these challenges, we propose strategies for leveraging AI-generated quizzes more effectively, such as incorporating targeted prompts and interactive workflows. Overall, this paper provides valuable insights and practical recommendations to enhance the alignment of AI tools with educational goals and improve the efficiency of quiz creation.

1 Introduction

Quizzes and assessments are fundamental in higher education, playing a crucial role in evaluating student understanding and shaping their learning experiences across diverse disciplines[1]. Whether in programming languages or artificial intelligence, these tools are cornerstone of courses that require students to balance foundational knowledge with practical application[2,3,4]. Beyond their evaluative purpose, quizzes encourage students to think critically and apply concepts in meaningful ways, fostering deeper engagement with the subject matter[5]. Traditionally, instructors dedicate significant time and effort to designing these assessments, ensuring they align closely with course objectives and effectively challenge learners. However, this process, while impactful, is both labor-intensive and time-consuming[6].

With the rise of Generative Artificial Intelligence (GenAI), the way we approach educational practices is changing[7]. Among these advancements is ChatGPT, a language model developed by OpenAI, capable of generating contextually relevant and coherent text by identifying patterns in data[8]. This technology has gained recognition for its ability to produce human-like content quickly and efficiently, opening up new possibilities for both students and educators[9].

In the educational landscape, ChatGPT presents significant opportunities for both students and instructors. For students, it functions as a personal tutor, offering instant feedback, answering questions, and fostering independent learning[10,11,12]. For educators, ChatGPT can assist with tasks such as lesson planning, providing detailed feedback, and notably, creating assessments[13,14]. By generating quiz questions tailored to specific topics, ChatGPT has the potential to save time and introduce a new level of customization to assessments.

This study investigates the effectiveness of using ChatGPT to create quizzes that align with learning objectives and accurately assess student understanding. It aims to provide evidence-based insights for educators and AI researchers by addressing the following research questions:

- **RQ1**: How to more effectively generate quizzes using AI?
- **RQ2**: How do AI-generated quizzes compare to manually crafted assessments in measuring understanding and differentiating performance?
- **RQ3**: How do students perceive the clarity, relevance, and overall quality of AI-generated quizzes compared to traditional ones?

To address these questions, we experimented with various prompts to generate quizzes using ChatGPT, refining the process to identify the most effective methods for creating quiz questions. We also collected student feedback through surveys and analyzed students' quiz results to compare the outcomes of AI-generated quizzes with manually generated quizzes. RQ1 is addressed in Section 3.2, while the answers of RQ2 and RQ3 are provided in Section 4. We further discuss the takeaways and future directions in Section 5. Through this exploration, the study provides insights into the practical use of generative AI in education.

The rest of this paper is organized as follows: Section 2 provides an overview of the background and related works on integrating GenAI into education. Section 3 discusses the methodology, including course organization, AI quiz generation, and the design of the comparative study with quantitative and qualitative analyses. Section 4 presents the results and key takeaways of our comparative study. Section 5 discusses the current limitations and future directions. Finally, Section 6 concludes the paper.

2 Background and Related Work

Assessments are a cornerstone of the educational process, enabling the evaluation of student understanding, the identification of knowledge gaps, and the adaptation of instructional strategies. These assessments are broadly categorized as formative and summative. Formative assessments, conducted during the learning process, provide ongoing feedback to help students improve and guide educators in refining their teaching methods[15]. Examples includes practice quizzes, in-class activities, and self-assessments. Summative assessments, in contrast, are designed to evaluate overall student learning at the end of a course, includes graded quizzes, exams, and projects[16]. Both types are crucial for achieving learning goals and ensuring that students meet course objectives[17,18].

The integration of GenAI into educational frameworks, particularly in computer science programs, is gaining popularity due to rapid technological advancements and the growing capabilities of these tools. Technologies like ChatGPT have become essential educational resources, offering immediate feedback, personalized tutoring, and practical assistance in coding. Studies highlight significant benefits, such as improved computational thinking, increased confidence and motivation among students in programming courses, underscoring the critical support these tools offer in computer science education[19,20]. However, challenges persist, including concerns about academic integrity, plagiarism risks, and potential impacts on critical thinking skills[21,22]. Faculty adoption of GenAI remains limited due to insufficient training and apprehensions about academic misconduct, emphasizing the need for targeted initiatives to equip educators and students with the skills to use these tools ethically and effectively[23,24,25,26]. Additionally, research highlights the importance of custom-tailored AI solutions that address diverse student needs, enhance engagement, and adapt teaching materials for better learning outcomes, particularly in online learning environments[27,28,20].

Recent studies have extensively explored the use of Generative AI (GenAI) for quiz generation, focusing on areas such as prompt engineering, integration design, and alignment with pedagogical frameworks[29,30,31]. These studies often evaluate the technical efficacy of GenAI-generated quizzes or their adherence to learning like Bloom's taxonomy, but they lack rigorous comparative testing in real educational settings[30].

In contrast, our study is the first to comparatively evaluate GenAI-generated and manually crafted quizzes in a live graduate-level course. By assessing these quizzes in terms of quality, relevance, alignment with course objectives, and student performance, this study valuable insights into the practical application of GenAI in education.

3 Methodology

This section introduces the methodology of our study. First, we outline the course background, followed by how we generated the quizzes using ChatGPT. Lastly, we introduce how the comparative study was conducted, including both the quantitative and qualitative analysis.

3.1 Course Organization and Learning Objectives

The study was conducted in a graduate-level AI course, which offers a comprehensive introduction to the foundational concepts, methods, and techniques of AI. The course is designed for students from diverse majors, including Software Engineering, Computer Engineering, Data Science, and Computer & Information Science. It covers a range of topics, such as knowledge representation methods, algorithms for reasoning, decision-making, planning, and learning, as well as modern intelligent systems capable of handling uncertainty. Through a combination of lectures, group activities, projects, and assessments, the course emphasizes both theoretical understanding and the practical application of AI concepts. Students engage in collaborative exercises and a project, allowing them to gain hands-on experience in analyzing and creating intelligent systems in real-world contexts.

3.2 Quiz Creation

3.2.1 Exploring the Effectiveness of Different Prompts

GenAI-based quiz creation involves prompt engineering. This process includes carefully crafting the instructions, or ”prompts,” provided to the AI system[32]. These prompts contain instructions and context to guide GenAI in generating questions that are clear, relevant, and aligned with the course content. By ensuring detailed and precise prompts, instructors can maximize the quality of the AI-generated quizzes.

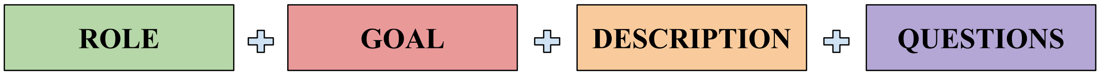

Figure 1: Prompt Formula

Our initial exploration follows the general formula[33] for prompt engineering, as shown in Figure 1. In this formula, “Role” refers to the specific style or persona for the AI agent to adopt. “Goal” is about setting a clear objective for the AI agent to achieve. “Description” includes additional details or steps for the AI agent to follow to achieve the goal. Lastly, “Questions” allow the AI agent to seek clarification on any uncertainties before responding. In particular, the initial prompt we used is given below:

Initial Prompt

Act as a professor teaching a graduate-level artificial intelligence, well-versed with all the AI topics **[Role]**. Your goal is to generate 15 multiple-choice quiz questions on the topics of machine learning, neural networks, and search algorithms that test conceptual understanding **[Goal]**. Each question should include four answer choices: one correct answer and three possible wrong answers. Ensure the question is clear, challenging, and aligns with graduate-level difficulty**[Description]**. Let me know if additional context or clarification is needed before generating the questions **[Questions]**.

3.2.2 Analyzing the Generated Quizzes

To assess the effectiveness of AI-generated quizzes, instructors and the teaching assistant (TA) analyzed and compared them with manually created quizzes. A total of 15 questions were included in both quizzes. The manual quizzes were created by instructors, while the AI-generated questions were produced by the teaching assistant (TA) using ChatGPT, which was provided with carefully designed prompts on the same topics. The topics covered in Quiz 1 and Quiz 2 are listed below:

1. **Quiz 1 Topics**: Agent Types (simple, reflex, goal-based, utility-based), Breadth-First Search, Depth-First Search, Big O Complexity, Greedy Search, Informed vs. Uninformed Search, Alpha-Beta Pruning, Genetic Algorithms (including key components such as population, crossover, and mutation), Confusion Matrix (True Positive, False Positive, True Negative, False Negative), Recall, Precision, and F1-Score.

2. **Quiz 2 Topics**: Strong AI vs. Weak AI, Technological Singularity, PEAS (Performance measure, Environment, Actuators, Sensors) for Agents, Agent Structure, Agent-Environment Interaction, Static vs. Dynamic Environments, Deterministic vs. Random Systems, Comparison of Breadth-First Search vs. Depth-First Search with examples, Completeness, Time and Space Complexity (Big O), Greedy Search vs. A*, Greedy Search vs. Uniform Cost Search, Depth-Limited Search vs. Iterative Deepening, Local Optima vs. Global Optima, Hill Climbing, Admissible Heuristics, Online vs. Offline Search, Perfect vs. Imperfect Games, Alan Turing and the Turing Test, Genetic Algorithm and its components, Logic and Propositional Inference, and evaluation metrics including True/False Positives, Recall, Precision, F1-Score, and the Confusion Matrix.

Example questions from Quiz 1 and Quiz 2 are provided in Table 1, respectively. We conducted a thorough manual analysis comparing AI-generated quiz questions with manually created ones, focusing on learning alignment, question relevance, and student performance. Each of the three co-authors independently reviewed all quiz questions, evaluating their alignment with learning objectives, clarity, and student performance based on past assessments. After completing their individual assessments, they discussed their observations and identified common themes that everyone agreed upon. Our key observations are summarized below.

	Manually Created Questions	**AI Generated Questions**
Q1	Generally, Breadth First Search performs better than Depth First Search if the tree is: (A) High branching and high depth (B) Low branching and high depth (C) Low branching and low depth (D) High branching and low depth Correct Answer: D	The time complexity of Breadth First Search is: (A) $O(n^2)$ (B) $O(b^d)$ (C) $O(d^b)$ (D) $O(n \log n)$ Correct Answer: C
Q2	Which of the following is a key difference between informed and uninformed search algorithms? (A) Informed search uses a heuristic function (B) Informed search does not use a goal (C) Uninformed search uses a heuristic function (D) Uninformed search is faster Correct Answer: A	Uniform Cost Search expands the node with the: (A) Lowest heuristic estimate. (B) Shortest path cost so far. (C) Highest priority value. (D) Maximum branching factor. Correct Answer: B

Table 1: Manual and AI Generated Quiz Questions Examples

1. **Out of Scope for the Course (Learning Alignment):** Some AI-generated quiz questions included topics beyond the intended course scope, such as large language models, which may misalign with the learning objectives.

2. **Uneven Distribution of Questions (Question Relevance):** Certain knowledge points have an excessive number of questions, resulting in potential overlap. Our analysis revealed that topics such as Breadth-First Search and Genetic Algorithms had a disproportionately high number of questions compared to other quiz topics. Additionally, we observed that some topics, including Perfect vs. Imperfect Games, Logic and Propositional Inference, and Precision, were not covered in the AI-generated quiz questions. This uneven distribution of questions effects the comprehensive understanding of students.

3. **Ambiguity in Questions (Student Performance**): We found that some AI-generated quiz questions on alpha-beta pruning and false negatives lacked sufficient context or detail, leading to ambiguity and making it difficult to determine the intended response. This ambiguity impacted students' ability to fully understand the questions. As a result, most students selected incorrect answers due to the lack of clarity and detail, ultimately affecting their performance and quiz scores.

3.2.3 Suggestions for Future Quiz Generation Prompts

To address these problems, we propose the following approaches to enhance the quality of the quizzes by leveraging additional context and incorporating the expertise of human:

- **Focus on Relevant Knowledge Points:** We used lecture notes and study guides to guide the generative AI (GenAI) in focusing on the relevant knowledge points. The additional prompt we used was: *"Focusing on the knowledge points in the attached document."* We also attempted to upload the lecture slides as additional context. However, this approach did not work as expected because the lecture slides covered a wide range of materials, and GenAI was unable to identify the focus by holistically analyzing the content. As a result, the questions generated from the lecture slides tended to emphasize details rather than the key knowledge points we intended to highlight.

- **Avoid Overlapping Questions:** To minimize redundancy, the additional prompt we used was: *"For each knowledge point, ask no more than one question."* For future quizzes with more questions, we recommend providing a detailed allocation of questions or points for each knowledge point.

- **Interactive Quiz Design:** Although we did not adopt this approach during quiz generation, we believe it would be effective to break the process into sequential steps. The first step would involve identifying which knowledge points require questions. After receiving confirmation from the instructor, GenAI could then more accurately generate quizzes aligned with the course learning objectives.

- **Generating Additional Questions:** Another potential solution is to semi-automatically generate quiz questions, allowing the human instructor to select the required number of questions from those generated. For example, we can ask GenAI to generate 50 questions, from which the instructor can manually select 15.

- **Avoid Ambiguity in Questions**: To ensure clarity in AI-generated quiz questions, instructors and TAs should carefully review them before use to address potential issues such as insufficient context or missing details that could lead to ambiguity.

However, since the goal was to evaluate the quizzes generated automatically by GenAI, the last two semi-automatic methods involving human input interactive quiz design and generating additional questions were not implemented during our quiz creation process.

3.3 Comparative Study: Quantitative and Qualitative Analysis

Using the above-mentioned quiz creation method, we designed a balanced study with two quizzes: one before the midterm and one after. Students were divided into two groups. For the first quiz, one group completed a manually created quiz, while the other took a quiz generated using ChatGPT. In the second quiz, the groups switched roles, ensuring that all students experienced both types of quizzes. This approach allowed us to compare the effectiveness of AI-generated quizzes with manually crafted ones while minimizing bias. The study employs a mixed-method approach, combining quantitative data from student performance with qualitative analysis through student interviews to gain deeper insights into the differences between the two types of quizzes.

3.3.1 Quantitative Analysis using Canvas Data

The students took the quizzes on Canvas, allowing us to easily leverage the platform's built-in analysis tools for data collection. This facilitated a detailed evaluation of quiz performance and question effectiveness. Specifically, we obtained the following metrics from the quiz statistics page on Canvas:

- **Quiz Summary:** This includes students' average score, highest score, lowest score, score distribution, standard deviation, and the average time taken by students to complete the quiz. These metrics provide an comprehensive understanding of the quiz's difficulty and the variability in student performance.
- **Discrimination Index:** This metric measures how well a quiz question distinguishes between high- and low-performing students. It is calculated as the difference between the proportion of top performers and low performers who answered correctly ($D = P_{\text{high}} - P_{\text{low}}$). Values range from -1 to +1, with higher values indicating better discrimination.
- **Difficulty Level:** Indicates the percentage of students who answered a question correctly ($\text{Difficulty Level} = \frac{\text{Correct Answers}}{\text{Total Attempts}} \times 100\%$). Lower percentages indicate more challenging questions, while higher percentages represents easier ones.

The quiz summary, discrimination index, and difficulty level collectively provide valuable insights into the effectiveness of quiz questions. The quiz summary offers a broad overview of the overall difficulty and student performance. The discrimination index evaluates how well questions differentiate between high-performing and low-performing students, highlighting the quality and alignment of questions with learning goals. The difficulty level measures the percentage of students who answered a question correctly, identifying questions that are either too easy or too challenging. Together, these metrics contribute to addressing RQ2: ”How do AI-generated quizzes compare to manually crafted assessments in measuring understanding and differentiating performance?”

3.3.2 Quantitative Analysis using Students’ Final Grades

We have the students’ final grades after finishing this course and use them to compare with the quiz results to determine whether the quiz results are aligned with students’ overall semester-long performance. To assess the degree of alignment between the quiz scores and the final grades, we use the Pearson correlation coefficient (r), which measures the strength and direction of the linear relationship between two continuous variables. The Pearson correlation coefficient is suitable when the data is continuous and normally distributed. Mathematically, r is defined as:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Where x_i and y_i are individual quiz scores and final grades, $\bar{x}$ and $\bar{y}$ are their means, and n is the total number of students. The value of r ranges from -1 to 1, where $r = +1$ indicates a strong positive alignment, meaning that as quiz scores increase, final grades also increase. An r value of 0 suggests no alignment, indicating no linear relationship between quiz scores and final grades. Conversely, $r = -1$ represents a strong negative alignment, where higher quiz scores are associated with lower final grades.

This analysis helps determine which quiz generation method more accurately reflects students’ overall learning performance, thereby addressing RQ2.

3.3.3 Quantitative Analysis using Surveys

	Survey Questionnaire
Q1	How relevant were the quiz questions to the topics covered in class?
Q2	How clear and understandable were the quiz questions?
Q3	How challenging did you find the quiz overall?
Q4	How confident are you in the accuracy and fairness of the quiz answers?
Q5	Did the quiz make you feel more prepared for the class topics?

Table 2: Survey Questionnaire

We designed a concise survey to capture student perceptions of key metrics. The survey consisted of five questions, as shown in Table 2, gathering students’ views on the relevance and clarity of the questions, their difficulty, their confidence in the accuracy and fairness of the answers, and their overall opinion on whether the quizzes were helpful in learning the class topics.

All 47 students in the class participated in the survey, and their responses were grouped according to the study design. Each student provided feedback after completing both an AI-generated quiz and a manually crafted quiz, ensuring balanced input from all participants. This approach provides valuable insights into RQ3: ”How do students perceive the clarity, relevance, and overall quality of AI-generated quizzes compared to traditional ones?”

3.3.4 Qualitative Analysis

In addition to the structured survey questions, we included an open-ended question to allow students to provide qualitative feedback on the quizzes. This approach aligns with qualitative research methodologies in educational studies, enabling us to capture nuanced student perspectives on the clarity, relevance, and overall quality of both AI-generated and manually crafted quizzes. These open-ended responses were analyzed to identify recurring themes and unique insights that complement the quantitative findings.

4 Results and Takeaways

This study evaluates the effectiveness of AI-generated quizzes compared to manually created ones by analyzing student scores, collecting survey data, and gathering qualitative feedback. The analysis focuses on relevance, clarity, difficulty, and overall student perceptions of both quiz formats.

4.1 Quantitative Analysis using Canvas Data

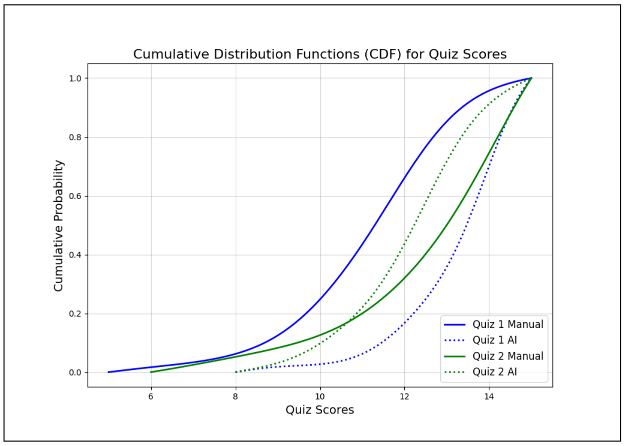

Figure 2: Quiz Score CDF Graph

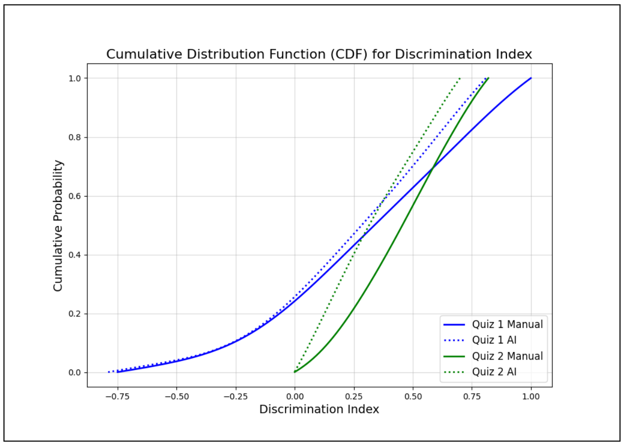

Figure 3: Discrimination Index CDF Graph

- **Quiz Summary**: The analysis of the CDFs for quiz scores shown in Figure 2, reveals differences between manually created and AI-generated quizzes. For Quiz 1, the manually created quiz allowed more students to achieve mid-range scores (8–11), while the AI-generated quiz was less challenging, with scores skewed toward the higher range. In Quiz 2, the manual quiz resulted in higher average scores overall, but the AI-generated quiz concentrated scores in the upper range (12–14). Overall, AI-generated quizzes tended to produce higher scores compared to manually created quizzes. This observation is further supported by the average time students took to complete the quizzes, with AI-generated quizzes requiring noticeably less time.

Quiz	Average Score	Standard Deviation	Max Score	Min Score	Average Time
Quiz 1 (Manual)	10.90	1.78	13.17	5.00	14:54 min
Quiz 1 (Gen AI)	13.17	1.44	14.00	8.00	10:35 min
Quiz 2 (Manual)	12.58	2.64	15.00	6.00	12:51 min
Quiz 2 (Gen AI)	11.08	1.29	13.00	7.00	11:00 min

Table 3: Students' Performance Statistics for Four Quizzes

Table 3 compares student performance on AI-generated and manually created quizzes. Both types of quizzes exhibits a similar pattern, with AI-generated quizzes showing a lower standard deviation in student scores.

- **Discrimination Index**: The analysis of the cumulative distribution function (CDF) for the discrimination index, shown in Figure 3 highlights a clear distinction between manually created and AI-generated quizzes. AI-generated quizzes consistently exhibited higher discrimination indexes, with approximately 70% of the questions achieving a score above 0.4, compared to 60% for manual quizzes. This indicates that AI-generated quizzes are generally more effective at differentiating between high- and low-performing students. Additionally, AI-generated quizzes demonstrated minimal variability, with a narrow range of discrimination indexes, reflecting consistency in question quality. In contrast, manual quizzes offered greater diversity, covering a broader range of discrimination indexes and including a few questions with lower scores in Quiz 1.
- **Difficulty Level**:

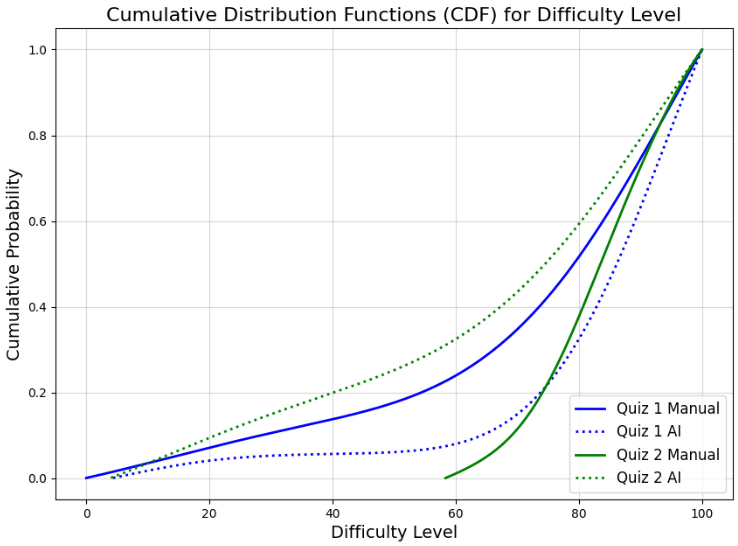

Figure 4: Difficulty Level CDF Graph

The CDF graph compares the difficulty levels of quizzes created manually and by generative AI. The difficulty index shows the percentage of students who answered correctly, with lower values indicating harder questions and higher values representing easier ones. Manual quizzes are more balanced, with many questions in the ideal difficulty range of 70-80%, which is effective for testing understanding. On the other hand, AI-generated quizzes have more very easy questions (>90%) and fewer challenging ones (<30%), as seen in their steeper curves. The difficulty level distribution explains the lower standard deviation in student scores for AI-generated quizzes.

Key Takeaways from Canvas Data Analysis

AI-generated quizzes result in higher student scores and less completion time compared to manually created quizzes, indicating they may be **easier** and more **straightforward**. However, **manual quizzes** exhibit a more **balanced difficulty distribution**, with fewer overly easy or overly difficult questions, and tend to challenge students more effectively, helping differentiate between high- and low-performing students.

4.2 Quantitative Analysis using Students' Final Grade

We compared the quiz results with the students' final grades to determine the degree of alignment between quiz scores and final grades. This was calculated using the Pearson correlation coefficient (r), and the values for both quizzes are provided in Table 4.

Quiz Type	**Quiz 1 Manual**	**Quiz 1 AI**	**Quiz 2 Manual**	**Quiz 2 AI**
Pearson Coefficient (r)	0.5	0.026	0.02	-0.353

Table 4: Pearson Correlation Coefficient (r) Values for Quizzes

The Pearson correlation coefficient between AI-generated quizzes and students' final scores is lower than that of manually created quizzes. Final scores are based on multiple assessment components, including discussion forums, midterm exams, quizzes, and final exams. Although quizzes contribute only a small portion of the final score, and students may excel in other assessments for various reasons, this finding suggests that AI-generated quizzes are less effective at reflecting students' overall learning performance and distinguishing between high- and low-performing students.

Key Takeaway from Final Grade Correlation Analysis

Manually created quizzes demonstrate a stronger correlation with students' final grades, indicating they better reflect overall learning and differentiate student performance, whereas AI-generated quizzes, while easier and more consistent, show weaker alignment with comprehensive course understanding.

4.3 Quantitative Analysis using Survey

All 47 students participated in the survey questionnaire. Questions Q1 to Q5 evaluated five key metrics: relevance, clarity, difficulty, accuracy and fairness, and overall usefulness in learning class topics. Students rated their opinions on a scale from 1 to 5, where 1 indicated the lowest rating (e.g., Not Relevant, Very Unclear, Very Easy, Very Doubtful, or Strongly Disagree), and 5 indicated the highest rating (e.g., Highly Relevant, Very Clear, Very Challenging, Very Confident, or Strongly Agree). The results, illustrated in the bar charts in Figure 5, reveal the differences in students' perceptions of manually created versus AI-generated quizzes.

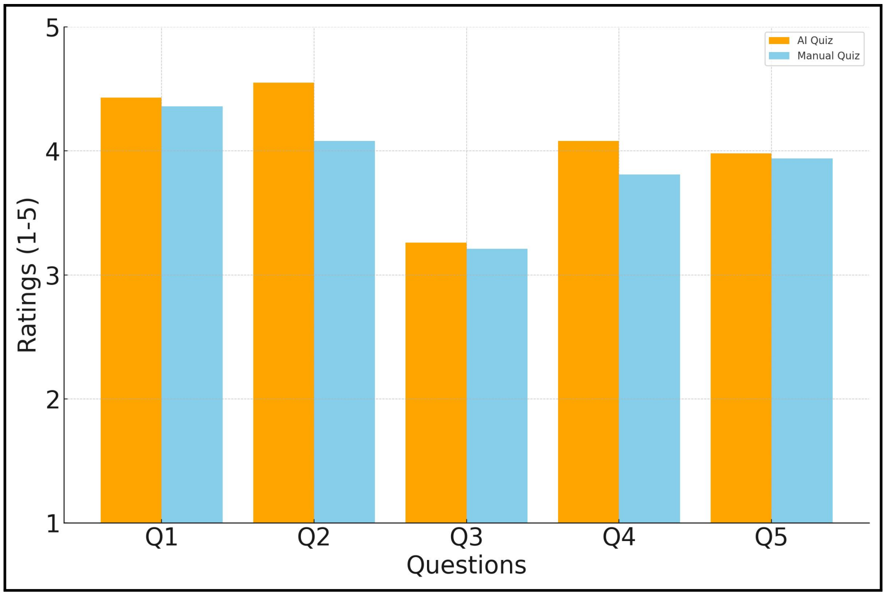

Figure 5: Comparison of AI-Generated vs Manually Created Quizzes

We observed significant differences in Q2 (Clarity) and Q4 (Accuracy and Fairness), where AI-generated quizzes received notably higher ratings than manually created quizzes. This finding helps explain the higher discrimination index for AI-generated quizzes, as students found the questions easier to understand, less ambiguous, and felt more confident in their answers.

Key Takeaway from Student Survey Analysis

Students found **AI-generated quizzes to be clearer and more consistent**, leading to higher ratings in clarity and fairness. However, **manually created quizzes were perceived as more challenging and engaging**, suggesting they may better support deeper learning and critical thinking.

4.4 Qualitative Analysis

To complement the quantitative findings, we conducted a qualitative analysis of student feedback using thematic coding to identify recurring patterns and insights. Open-ended survey responses were systematically analyzed to categorize key themes, such as clarity, difficulty, relevance, and engagement. Below we select some comments from students:

Manually created quizzes: *"I appreciated how the questions were tailored to specific topics we covered in class, allowing me to connect concepts better.", "The questions were clear for the most part, but some of the answers were a little tricky with how they were worded which made simple questions a little more challenging.", " Some of the questions felt tricky and sounded that that they can be interpreted in different ways."*

AI generated quizzes: *“Quiz questions were not that much hard. Yes they are fully relevant to the classes.“, All of the questions were straightforward and it was easy.“, “It helped refreshing the topics covered during the lectures.“*

Student feedback highlights the unique strengths of both AI-generated and manually created quizzes. Manual quizzes were praised for their tailored approach to course material, fostering deeper engagement and critical thinking, although some students found them occasionally tricky or open to interpretation. On the other hand, AI-generated quizzes excelled in clarity and consistency, aligning with our observations from the discrimination index and survey. While manual quizzes offer creative depth, AI-generated quizzes provide structure, suggesting that combining these approaches could create a balanced assessment strategy.

Key Takeaway from Qualitative Analysis

Student feedback highlights the unique strengths of both quiz formats. **Manually created quizzes were praised for their alignment with course material and their ability to foster critical thinking,** though some students found them tricky or ambiguous. In contrast, **AI-generated quizzes were valued for their clarity and consistency,** making them easier to understand but potentially less challenging. These insights suggest that combining both approaches could create a more balanced assessment strategy.

4.5 Summarizing Key Takeaways

Summarizing the results, this study identifies the key strengths and limitations of AI-generated quizzes compared to manually created ones, based on student performance, survey data, and qualitative feedback:

- **Strengths of AI-generated Quizzes:** AI-generated quizzes excel in clarity and consistency, as highlighted by higher student ratings in Q2 (Clarity) and Q4 (Accuracy and Fairness), as well as a higher discrimination index. These quizzes provide straightforward and accessible questions that effectively assess basic understanding, making them particularly well-suited for self-assessment.
- **Limitations of AI-generated Quizzes:** Despite their clarity, AI-generated quizzes tend to skew toward easier questions, resulting in a less balanced distribution of difficulty levels. This results in a lower standard deviation in scores and a reduced ability to differentiate high-performing students from low-performing ones. The weaker correlation between AI-generated quiz scores and final grades further suggests that these quizzes may not fully reflect students’ overall learning performance.

5 Discussion and Future Directions

Generative AI, by its nature, operates as a static, knowledge-based approach that reflects information already created by humans. Previous studies have explored the role of generative AI in quiz creation, with mixed results. Some research suggests that AI-generated quizzes improve efficiency and provide high-quality questions with clear phrasing and structure[29]. Our findings align with these studies, as students rated AI-generated quizzes higher in clarity and fairness.

However, our results also indicate that AI-generated quizzes tend to be easier and less effective in differentiating student performance, which contrasts with prior work emphasizing AI's potential to generate adaptive or personalized assessments[31]. This suggests that while AI-generated quizzes can streamline assessment creation, they may require manual refinement to enhance difficulty and depth. In contrast, manually crafted quizzes demonstrated better alignment with final grades, supporting research that highlights the pedagogical value of instructor-designed assessments in fostering critical thinking and deeper engagement[4,16]. Our results reinforce these conclusions by showing that manual quizzes were more effective at differentiating high- and low-performing students, even though they required significantly more time to design. This suggests that AI-generated quizzes, while efficient, may not yet be a substitute for human expertise in assessment design.

As discussed in the previous sections, we envision that **semi-automated quiz generation** can address these challenges. The first strategy involves breaking the process into sequential steps, identifying the key knowledge points requiring questions, followed by instructor confirmation to ensure alignment with course objectives. The second strategy proposes generating a larger pool of quiz questions, such as 50, from which instructors can manually select the most relevant 15. These approaches highlight the importance of human oversight in enhancing the relevance and quality of quiz content, while also leveraging GenAI's efficiency. Furthermore, this collaborative process not only improves immediate outcomes but also generates valuable training data that can further refine GenAI's performance over time.

That being said, GenAI holds significant potential for addressing common challenges in education. For example, one persistent issue many instructors face is the availability of past exams online, which enables some students to achieve high grades without truly mastering the material. This creates unfairness in assessment and undermines the learning process. By leveraging past quizzes and guiding GenAI to intelligently modify the questions, instructors can easily generate new and diverse assessments that minimize this risk and maintain fairness.

Beyond quiz generation, GenAI offers opportunities to derive deeper insights from assessment results. For instance, it can analyze quiz outcomes to identify students' weak points, providing instructors with data-driven insights to tailor their teaching strategies. Moreover, GenAI can support personalized learning by suggesting customized quizzes or learning materials to help individual students address their knowledge gaps. This opens avenues for a more adaptive and student-centered educational experience, further enhancing the role of GenAI as a valuable tool in modern classrooms.

Another promising opportunity lies in integrating GenAI directly into educational platforms like Canvas. By leveraging GenAI's capabilities, Canvas could offer built-in tools to automatically generate quiz banks based on instructors' uploaded materials, such as lecture notes, slides, and course outlines. This integration would streamline the quiz creation process, enabling students to quickly access a variety of questions aligned with their course objectives for self-assessment. Additionally, the platform could provide options for instructors to review, edit, and refine the questions directly within Canvas, ensuring that the final assessments maintain quality and relevance.

While our study provides evidence-based insights into the rapidly evolving field of

GenAI-supported education, it is important to acknowledge its limitations. A key limitation is the relatively small sample size, which may not fully reflect the variability in learning outcomes across diverse student populations. Furthermore, the study was conducted within the context of a single course, which limits the generalizability of our findings to other courses, subjects, or educational settings. Future studies should explore AI-generated quizzes across diverse subjects, undergraduate courses, and different AI models to validate these findings. Additionally, future research could examine long-term learning outcomes, including whether AI-generated quizzes help with knowledge retention and concept mastery.

6 Conclusion

This study explores the use of Generative AI, specifically ChatGPT, to automating quiz creation in a graduate-level AI course. Through a comparison of AI-generated and manually created quizzes, the research highlights the unique strengths and limitations of each approach. AI-generated quizzes excel in clarity, consistency, and efficiency, making them a practical tool for streamlining assessment creation. However, manual quizzes offer greater flexibility and creativity, fostering deeper engagement and critical thinking, though they require more time and effort to design. The findings provide valuable insights into how Generative AI can support educators in creating assessments aligned with course objectives.

References

[1] T. Baumeister, P. Rambach, and D. Fey, "The benefits of continuous assessment: A case study on the effectiveness of weekly online quizzes in computer science courses," in *ICERI2023 Proceedings*, pp. 4658–4664, IATED, 2023.

[2] K. E. Dunn and S. W. Mulvenon, "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*, vol. 14, no. 1, p. 7, 2019.

[3] K. Martens, D. Niemann, and J. Teltemann, "Effects of international assessments in education–a multidisciplinary review," *European Educational Research Journal*, vol. 15, no. 5, pp. 516–522, 2016.

[4] D. P. Collins, D. Rasco, and V. A. Benassi, "Test-enhanced learning: Does deeper processing on quizzes benefit exam performance?," *Teaching of Psychology*, vol. 45, no. 3, pp. 235–238, 2018.

[5] B. W. Kennedy, "The value of quizzes," *Lab Animal*, vol. 44, no. 10, pp. 409–409, 2015.

[6] B. Khan, S. Chenda, S. Heng, and D. Coniam, ""doing a quiz in pyjamas": Successes and challenges of blended learning in cambodian higher english language education," *Blended learning for inclusive and quality Higher Education in Asia*, pp. 125–150, 2021.

[7] D. Wood and S. H. Moss, "Evaluating the impact of students' generative ai use in educational contexts," *Journal of Research in Innovative Teaching & Learning*, vol. 17, no. 2, pp. 152–167, 2024.

[8] K. I. Roumeliotis and N. D. Tselikas, "Chatgpt and open-ai models: A preliminary review," *Future Internet*, vol. 15, no. 6, p. 192, 2023.

[9] S. Feuerriegel, J. Hartmann, C. Janiesch, and P. Zschech, “Generative ai,” *Business & Information Systems Engineering*, vol. 66, no. 1, pp. 111–126, 2024.

[10] D. Baidoo-Anu and L. O. Ansah, “Education in the era of generative artificial intelligence (ai): Understanding the potential benefits of chatgpt in promoting teaching and learning,” *Journal of AI*, vol. 7, no. 1, pp. 52–62, 2023.

[11] T. Rasul, S. Nair, D. Kalendra, M. Robin, F. de Oliveira Santini, W. J. Ladeira, M. Sun, I. Day, R. A. Rather, and L. Heathcote, “The role of chatgpt in higher education: Benefits, challenges, and future research directions,” *Journal of Applied Learning and Teaching*, vol. 6, no. 1, pp. 41–56, 2023.

[12] S. Sok and K. Heng, “Chatgpt for education and research: A review of benefits and risks,” *Cambodian Journal of Educational Research*, vol. 3, no. 1, pp. 110–121, 2023.

[13] A. ElSayary, “An investigation of teachers’ perceptions of using chatgpt as a supporting tool for teaching and learning in the digital era,” *Journal of computer assisted learning*, vol. 40, no. 3, pp. 931–945, 2024.

[14] J. Whalen, C. Mouza, *et al.*, “Chatgpt: challenges, opportunities, and implications for teacher education,” *Contemporary Issues in Technology and Teacher Education*, vol. 23, no. 1, pp. 1–23, 2023.

[15] R. E. Bennett, “Formative assessment: A critical review,” *Assessment in education: principles, policy & practice*, vol. 18, no. 1, pp. 5–25, 2011.

[16] P. T. Knight, “Summative assessment in higher education: practices in disarray,” *Studies in higher Education*, vol. 27, no. 3, pp. 275–286, 2002.

[17] E. B. Nuhfer, “The place of formative evaluations in assessment and ways to reap their benefits,” *Journal of Geoscience Education*, vol. 44, no. 4, pp. 385–394, 1996.

[18] F. Fitriani, “Implementing authentic assessment of curriculum 2013: Teacher’s problems and solusions,” *Getsempena English Education Journal*, vol. 4, no. 2, 2017.

[19] R. Yilmaz and F. G. K. Yilmaz, “The effect of generative artificial intelligence (ai)-based tool use on students’ computational thinking skills, programming self-efficacy and motivation,” *Computers and Education: Artificial Intelligence*, vol. 4, p. 100147, 2023.

[20] V. Roger-Monzó, “Impact of generative artificial intelligence in higher education: Student perceptions,” in *INTED2024 Proceedings*, pp. 2631–2635, IATED, 2024.

[21] R. Michel-Villarreal, E. Vilalta-Perdomo, D. E. Salinas-Navarro, R. Thierry-Aguilera, and F. S. Gerardou, “Challenges and opportunities of generative ai for higher education as explained by chatgpt,” *Education Sciences*, vol. 13, no. 9, p. 856, 2023.

[22] G. Cooper, “Examining science education in chatgpt: An exploratory study of generative artificial intelligence,” *Journal of Science Education and Technology*, vol. 32, no. 3, pp. 444–452, 2023.

[23] B. Obenza, A. Salvahan, A. N. Rios, A. Solo, R. A. Alburo, and R. J. Gabila, “University students’ perception and use of chatgpt: Generative artificial intelligence (ai) in higher education,” *International Journal of Human Computing Studies*, vol. 5, no. 12, pp. 5–18, 2024.

[24] A. Kelly, M. Sullivan, and K. Strampel, “Generative artificial intelligence: University student awareness, experience, and confidence in use across disciplines,” 2023.

[25] C. Bopp, A. Foerst, and B. Kellogg, “The case for llm workshops,” in *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*, pp. 130–136, 2024.

[26] V. A. Kusam, S. Shrestha, K. Kattan, B. Maxim, and Z. Song, “A pbl-based mini course module for teaching computer science students to utilize generative ai for enhanced learning,” in *2024 IEEE Frontiers in Education Conference (FIE)*, pp. 1–9, IEEE, 2024.

[27] L. I. Ruiz-Rojas, P. Acosta-Vargas, J. De-Moreta-Llovet, and M. Gonzalez-Rodriguez, "Empowering education with generative artificial intelligence tools: Approach with an instructional design matrix," *Sustainability*, vol. 15, no. 15, p. 11524, 2023.

[28] A. Smolansky, A. Cram, C. Raduescu, S. Zeivots, E. Huber, and R. F. Kizilcec, "Educator and student perspectives on the impact of generative ai on assessments in higher education," in *Proceedings of the tenth ACM conference on Learning@ Scale*, pp. 378–382, 2023.

[29] S. Hutt and G. Hieb, "Scaling up mastery learning with generative ai: Exploring how generative ai can assist in the generation and evaluation of mastery quiz questions," in *Proceedings of the Eleventh ACM Conference on Learning@ Scale (L@S '24)*, 2024.

[30] S. Elkins, E. Kochmar, J. C. Cheung, and I. Serban, "How teachers can use large language models and bloom's taxonomy to create educational quizzes," in *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI-24)*, 2024.

[31] J. Hassell, "Best practices for using generative ai to create quiz content for the canvas lms," in *2024 ASEE Midwest Section Conference*, 2024.

[32] D. C. Schmidt, J. Spencer-Smith, Q. Fu, and J. White, "Towards a catalog of prompt patterns to enhance the discipline of prompt engineering," 2023.

[33] V. A. Kusam, L. Moore, S. Shrestha, Z. Song, J. Lu, and Q. Zhu, "Generative-ai assisted feedback provisioning for project-based learning in cs courses," in *2024 ASEE Annual Conference & Exposition*, 2024.