A Cross-Disciplinary Study Evaluating the Effectiveness of GenAI created Multiple Choice Questions

Erica Perich, zyBooks, A Wiley Brand

Erica Perich is Authoring and Pedagogy lead at zyBooks. She works on incorporating learning science into pedagogical practices for authoring across disciplines. She has also extensively contributed as an author to a leading online IT digital textbook. She earned a BS in Mathematics Education from Brigham Young University in 2013 and an MS in Information Technology from Southern New Hampshire University in 2023. While teaching secondary mathematics, she gained a passion for finding innovative ways to facilitate students' conceptual understanding.

Dr. Yamuna Rajasekhar, zyBooks, A Wiley Brand

Yamuna Rajasekhar is a senior manager of Content at zyBooks, a Wiley Brand. She is an author and contributor to various zyBooks titles. She was formerly an assistant professor of Electrical and Computer Engineering at Miami University. She received her M.S. and Ph.D. in Electrical and Computer Engineering from UNC Charlotte.

A Cross-Disciplinary Study Evaluating the Effectiveness of GenAI-Created Multiple Choice Questions

As GenAI tools become increasingly available in education, instructors are actively exploring how to leverage AI to manage growing workloads, particularly in assessment design, where creating and maintaining question banks for homework, quizzes, and exams requires considerable time and effort. However, the effectiveness of AI-generated assessments remains an open question, with instructors uncertain about question quality and reliability of the AI system.

Multiple choice questions (MCQs) represent a particularly compelling use case for AI automation given their widespread use, objectivity, and need for large question banks. However, the use of poorly engineered prompts can negate the advantages of using AI by generating questions that are flawed or irrelevant, requiring more effort to evaluate and often need extensive edits. Systematic frameworks for evaluating AI-generated MCQ effectiveness and empirical evidence of GenAI's adherence to quality guidelines remain limited.

We introduce an AI-powered MCQ generation tool within an online interactive textbook platform to generate assessments. The tool uses GPT-4.1 with structured prompts that incorporate explicit guardrails and learning science principles to generate high quality MCQs from instructor-selected content.

To evaluate whether GenAI reliably adheres to these prompt-based guardrails, we present a multi-dimensional framework and apply it to over 500 questions generated for introductory courses in Computer Science, Mathematics, and Data Science. Our framework evaluates the efficacy of MCQs based on quality and relevance, each of which are defined by several quantifiable metrics. Overall, our findings reveal the extent to which GenAI adheres to prompt-based guardrails and generates effective MCQs. We present both strengths and systematic failure patterns that inform best practices for GenAI-assisted assessment design.

Introduction

The proliferation of generative AI in educational contexts has prompted both instructors and students to explore diverse applications of these technologies. Students primarily utilize GenAI tools for academic support, including study assistance and assignment completion, with AI-enabled tutoring systems representing the most prevalent integration model within learning platforms. However, the deployment of student-facing AI tools introduces significant pedagogical concerns, particularly regarding the propagation of AI-generated inaccuracies. Students at varying stages of content mastery may lack the expertise to identify when AI presents erroneous information, potentially undermining learning outcomes and reinforcing misconceptions.

Despite these risks, GenAI offers substantial advantages that continue to advance rapidly. Rather than restricting AI integration entirely, a more measured approach involves prioritizing instructor-facing applications, where domain experts can validate AI-generated content before student exposure. This strategic deployment addresses a critical challenge in higher education: the persistent scarcity of instructor time for course preparation. Given that instructors dedicate considerable effort to developing learning materials and assessments outside the classroom, AI assistance in these domains represents a high-impact opportunity for workload management without compromising educational quality.

Multiple choice questions exemplify an assessment format where AI assistance could provide substantial value. MCQs are widely employed across quizzes, examinations, and formative assessments, requiring both extensive question banks and meticulous construction. Effective MCQ development demands carefully crafted question stems paired with plausible distractors that ideally reflect common student misconceptions. The assessment format's efficiency, enabling rapid evaluation of student understanding, necessitates large question volumes, with typical examinations requiring dozens of items. This volume requirement creates a significant time burden for instructors. The present study addresses this challenge by introducing and evaluating an AI-powered MCQ generation tool specifically designed to produce high-quality assessment items aligned with pedagogical best practices.

Background

In recent years, there have been several explorations of AI usage in the classroom and in different learning materials. One notable example is the experiment of Google's LearnLM for tutoring. This paper [1] shows that LearnLM proved to be highly reliable with human intervention where expert human tutors approved 76% of messages without edits and finding harmful content. Students that received AI tutoring showed significantly better knowledge transfer to new topics and performed just as well as those with human tutors. The overall findings show that AI tutoring can effectively scale personalized learning support when supervised.

Satheesh et al [2] presents an AI-powered system that generates MCQs using Natural Language Processing (NLP) techniques including named entity recognition and part-of-speech recognition. The system is developed to process multiple formats including PDF, DOCX, PPTX, and TXT. The system rapidly generates quizzes in seconds by using a structured workflow of document processing combined with NLP analysis, and question formulation with automatic distractor generation and grammar checking. Educators reported a 40% reduction in preparation time and students reporting that the quizzes are effective for self assessment. A key challenge reported was a lack of adaptive difficulty adjustment based on learner performance.

While AI-powered tools offer promise for automating MCQ generation at scale, evaluating the quality of both human-authored and AI-generated questions requires validated item-writing standards. Haladyna, Downing, and Rodriguez [3] validated 31 multiple-choice item-writing guidelines by synthesizing evidence from 27 educational measurement textbooks and 27 empirical studies. Their taxonomy organizes guidelines into five categories: Content Concerns, Formatting Concerns, Style Concerns, Writing the Stem, and Writing the Choices. Key validated guidelines include ensuring items test important content with clear stems, using simple vocabulary, developing plausible distractors, and avoiding cues to correct answers. The authors found that guidelines like "none of the above" and negatively worded stems require cautious use due to their effects on item difficulty and differential impact on student ability levels.

Building on these guidelines, Breakall, Randles, and Tasker [4] developed the Item Writing Flaws Evaluation Instrument (IWFEI) to operationalize item-writing flaw detection in the context of general chemistry education. The IWFEI demonstrated high inter-rater reliability (91.8% agreement, Krippendorff's Alpha = 0.836) and was applied to analyze 1,019 general chemistry MCQs. The study found that 83% of items contained at least one item-writing flaw, with implausible distractors being the most common issue. This work demonstrated that systematic evaluation of item-writing flaws is feasible and revealed widespread quality issues even in instructor-created assessments.

Moore, Costello, Nguyen, and Stamper [5] introduced SAQUET (Scalable Automatic Question Usability Evaluation Toolkit), an open-source tool that automates MCQ quality evaluation using the Item-Writing Flaws (IWF) rubric. SAQUET leverages GPT-4, word embeddings, and Transformers to identify design flaws in MCQs, addressing limitations of traditional automated evaluation methods that prioritize readability while overlooking deeper structural issues. The authors demonstrated a discrepancy between commonly used automated metrics and human quality assessments, then validated SAQUET across five domains (Chemistry, Statistics, Computer Science, Humanities, and Healthcare), showing that it effectively distinguishes between flawed and well-constructed questions.

Methods

We evaluate the efficacy of multiple choice questions using a multi-dimensional framework consisting of quality and relevance. Quality measures whether a question is well-constructed according to established best practices in educational assessment, while relevance measures whether a question assesses the intended learning objectives for the associated material.

Quality evaluation

We define MCQ quality by three core characteristics: clarity, accuracy, and test-wiseness prevention. In order to be an effective multiple choice question, an MCQ must be clear to students, technically correct, and a true assessment.

Clarity includes both conceptual and linguistic dimensions. **Conceptual clarity** ensures an MCQ is focused on the essential aspects of the topic. An MCQ lacking conceptual clarity risks assessing students on an unrelated topic or tangential knowledge rather than the intended learning objective, making the question an ineffective assessment. A conceptually clear multiple choice question focuses on a single, core concept; assesses fundamental understanding; and avoids reliance on trickery or trivial detail. **Linguistic clarity** ensures an MCQ is easily comprehensible to students. An MCQ lacking linguistic clarity can lead to unnecessary student confusion. This linguistic ambiguity can negatively impact the performance of students who actually understand the concept being assessed, rendering the question ineffective. A linguistically clear multiple choice question employs simple, concise language and uses consistent terminology to describe technical terms and concepts.

Accuracy requires that an MCQ is technically correct and has exactly one defensible correct answer. An inaccurate MCQ fails to reliably measure student understanding. Technical errors in the question or answer choices may result in multiple correct answers or no correct answer at all, penalizing students with accurate knowledge and making the assessment invalid.

Test-wiseness prevention provides assurance that an MCQ does not contain inadvertent cues that allow students to select the correct answer through test-taking strategies rather than content knowledge. An MCQ that lacks test-wiseness prevention advantages students who are skilled in recognizing patterns in MCQ construction, regardless of their understanding of the content. This undermines the validity of the MCQ by assessing test-taking skills rather than the intended learning objective. An MCQ prevents test-wiseness by avoiding systematic patterns in correct answer placement or length, eliminating word overlap between the question stem and correct answer, and avoiding options that allow for partial-knowledge guessing.

We evaluate the quality of both AI-generated and subject matter expert authored MCQs using the Scalable Automatic Question Usability Evaluation Toolkit (SAQUET) [5], an AI-powered evaluation tool based on the **Item Writing Flaw (IWF)** rubric. SAQUET returns a pass/fail result for each of the item writing flaws in the IWF rubric. Based on prior research [5], questions with 0 or 1 criteria failure are considered acceptable, while questions with 2 or more failures are considered unacceptable.

We map each IWF criterion to one of our three quality dimensions based on the type of flaw it detects (Table 1). Clarity criteria identify issues that impede comprehension or focus, accuracy criteria detect technical incorrectness, and test-wiseness prevention criteria flag cues that allow students to deduce the correct answer without adequate understanding of the concept.

Quality dimension	IWF criteria
Clarity	● Ambiguous information ● Complex or K-type ● Negatively worded ● Lost sequence ● Unfocused stem ● Vague terms ● Gratuitous information
Accuracy	● Absolute terms ● More than one correct
Test-wiseness prevention	● Longest option correct ● Implausible distractors ● True or false ● Convergence cues ● None of the above ● Word repeats ● Logical cues ● All of the above ● Grammatical cues

Table 1: Distribution of IWF criteria across quality dimensions

Quality metrics

- Question acceptance rate: The percentage of questions classified as acceptable (0 or 1 failure) versus unacceptable (2 or more failures) for both AI-generated and subject matter expert authored questions, overall and by discipline.
- Failure distribution: The distribution showing the count of questions with 0, 1, 2, 3, or more criterion failures, showing the overall quality profile.
- Failure rate by dimension: The proportion of questions failing criteria within each of the three dimensions (Clarity, Accuracy, Test-wiseness prevention), enabling identification of specific quality strengths and weaknesses.
- Failures by criterion: The failure rate for each individual IWF criterion, highlighting which specific guidelines are most frequently violated.

Relevance assessment

We define relevance as alignment with section learning objectives (LOs). A high-quality MCQ is only effective if it assesses students on the intended learning objective. Automated relevance measurement was conducted using OpenAI's GPT-4.1-mini model. The model was first asked to infer three to five learning objectives given the section content. Then, for each MCQ, the model was instructed to determine if the question directly assesses one or more of the inferred learning objectives and return a yes/no response with a one sentence justification.

Relevance metrics

- LO alignment rate: The percentage of questions classified as aligned versus not aligned, overall and by discipline.

Sample selection

We selected 60 sections across four zyBooks, covering introductory courses for three disciplines: two computer science zyBooks (Programming in Java and Programming in Python 3), one data science zyBook (Data Science Foundations with Python), and one mathematics zyBook (Discrete Mathematics). For each zyBook, we selected the 15 most-used sections, determined by the number of distinct users, in order to have the sample comprise the most commonly taught topics in introductory courses across the three disciplines.

Question generation strategy

We generated multiple choice questions using a custom MCQ generation tool, powered by OpenAI's GPT-4.1-mini. The number of questions generated for each section was directly correlated with the number of subject matter expert authored questions already existing for that section. This approach was used because the existing number of subject matter expert authored questions is considered a reliable measure of the necessary content coverage for the topic. The same prompt was run four times for each section to produce four sets of questions for each of the 60 sections.

To enable cross-run consistency analysis across disciplines, we conducted additional generation for the top five sections in each zyBook (20 sections total). For each of these sections, we generated 4 question sets of 10 questions each, providing standardized question counts for direct comparison of consistency metrics across disciplines.

In total, 2,052 AI-generated MCQs were generated across three disciplines: 1,252 for efficacy analysis across 60 topics and an additional 800 for cross-disciplinary consistency analysis.

Fig. 1 shows the question generation process. The user first inserts section content and initiates question generation (Fig. 1(a)). The tool then processes the request and generates questions using GPT-4.1-mini (Fig. 1(b)). Finally, the generated questions are shown in the content preview pane (Fig. 1(c)).

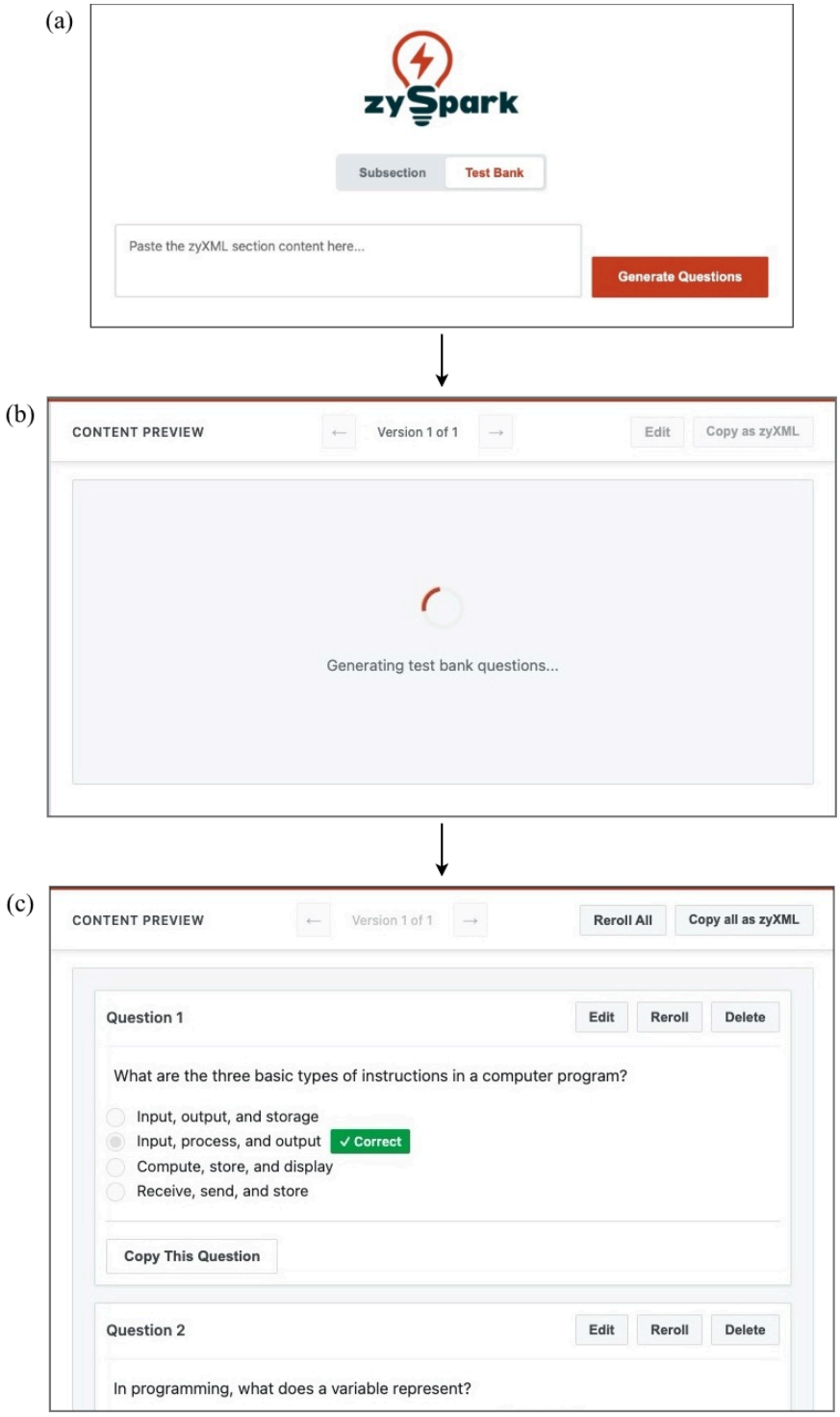

Figure 1: Workflow of the custom MCQ generation tool.

A representative sample of questions generated by this process for each discipline is provided in Appendix A.

During question generation, the model was provided with a carefully crafted prompt alongside the section content. This prompt operationalized our efficacy framework though explicit constraints targeting quality and relevance. Key prompt constraints are shown below (verbatim from the prompt provided to the model).

Quality constraints

- Use simple, clear language (prefer "use" over "utilize")
- Each question addresses only ONE concept
- Keep questions concise: 15-20 words
- Ensure each question is technically correct
- Answer choices should be in the same form (parallel)
- Avoid absolute words: only, never, always

Relevance constraints

- Questions should focus on most important concepts
- Focus on core understanding, not edge cases (unless central to concept)

Providing these explicit constraints aligned MCQ generation with our efficacy framework and enabled systematic examination of the model's adherence to quality and relevance requirements.

Subject matter expert authored comparison set

We conducted quality analysis on subject matter expert authored MCQs for the same 60 sections to serve as a quality benchmark for the AI-generated MCQs. These questions were originally authored by subject matter experts to serve as test bank questions that assess students on the section's critical content. A total of 313 subject matter expert authored questions were analyzed across the three disciplines.

Consistency and duplication analysis

To assess generation consistency and identify redundant content, we analyzed semantic similarity across questions using a custom analysis pipeline. We concatenated each question's stem with its answer choices and processed the text through OpenAI's text-embedding-3-small model to generate vector embeddings. For each set of questions, we computed pairwise cosine similarity for all $N \times (N-1)/2$ question pairs.

We classified relationships using two thresholds: question pairs with cosine similarity ≥ 0.85 were classified as duplicates, while question pairs with cosine similarity between 0.70 and 0.85 were classified as similar but distinct.

Results and discussion

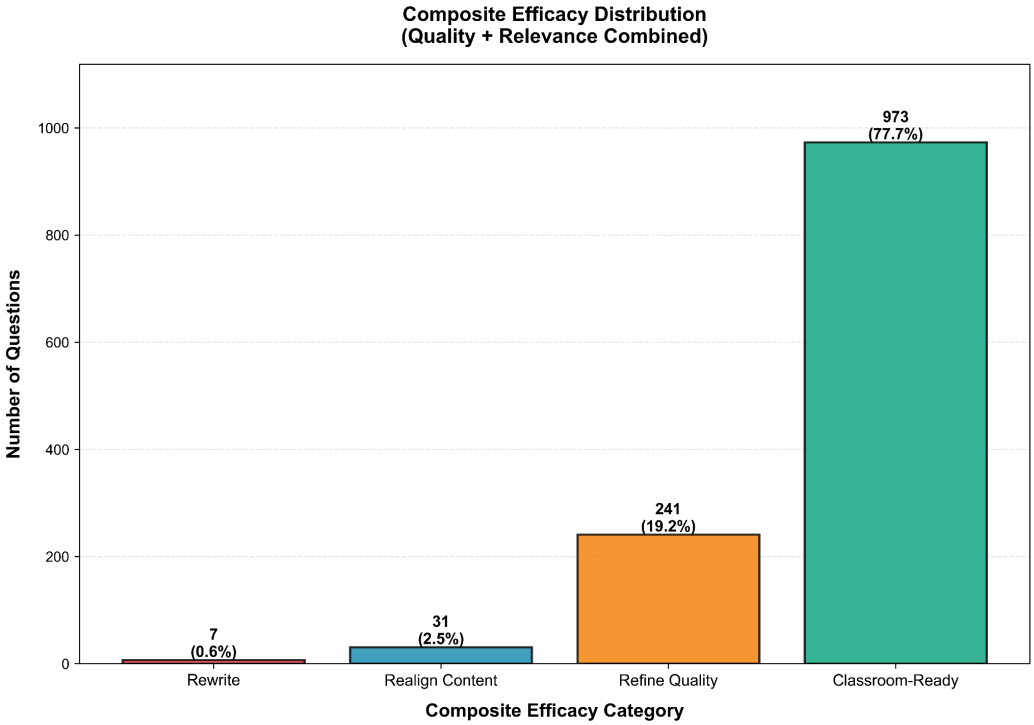

Figure 2: Composite efficacy distribution

Fig. 2 shows out of the 1,252 AI-generated MCQs, 973 (77.7%) were found to be classroom ready, meaning they were both high quality and aligned to the content's learning objectives. A further 241 (19.2%) were aligned to the learning objectives, but would require quality refinement for classroom use, which might include factors like ambiguous information and implausible distractors. Around 31 questions required realignment to learning objectives. Only 7 questions (0.6%) required a full rewrite, which is an encouraging preliminary result. An example of each outcome is provided in Appendix B.

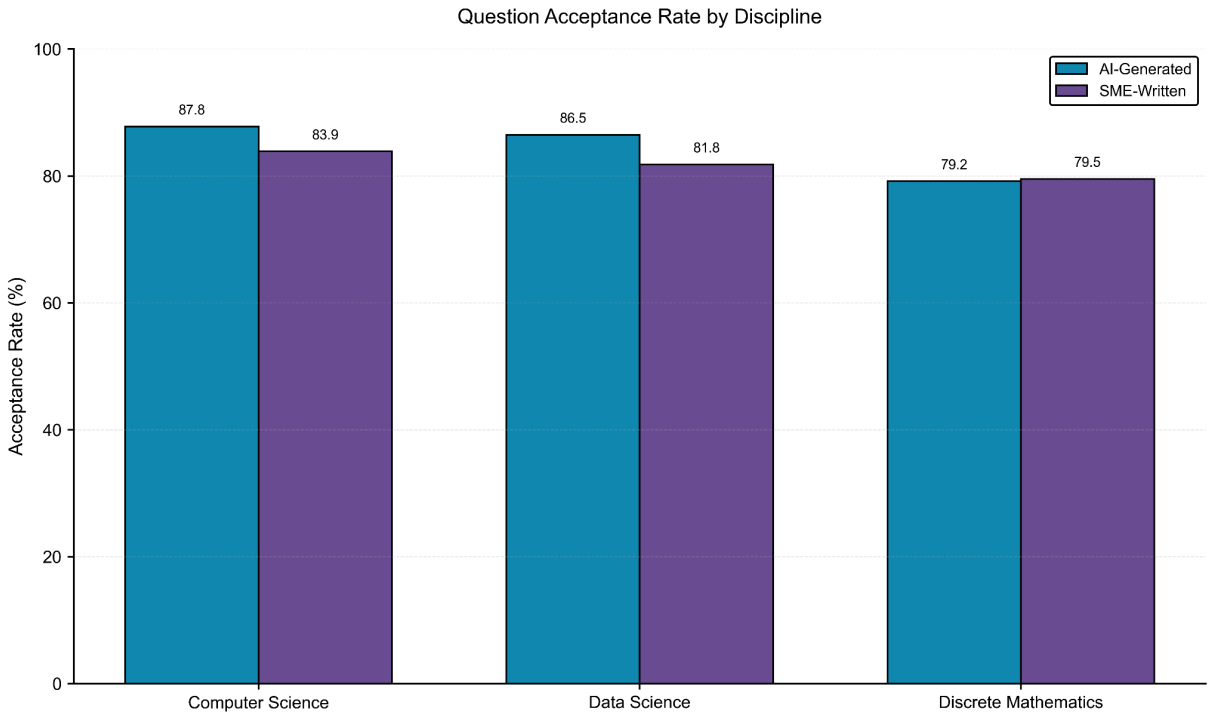

Figure 3: Question acceptance rate by discipline.

The question acceptance rate for computer science was 87.8% for the AI generated questions and 83.9 for SME-authored questions. While this is very positive, a key aspect to consider here is the time savings to author the questions. The average time to author one quality question by a SME is about 20 minutes while the AI tool took about 0.9 seconds per question. Even with a SME in the loop to approve questions, AI generated questions offer large savings.

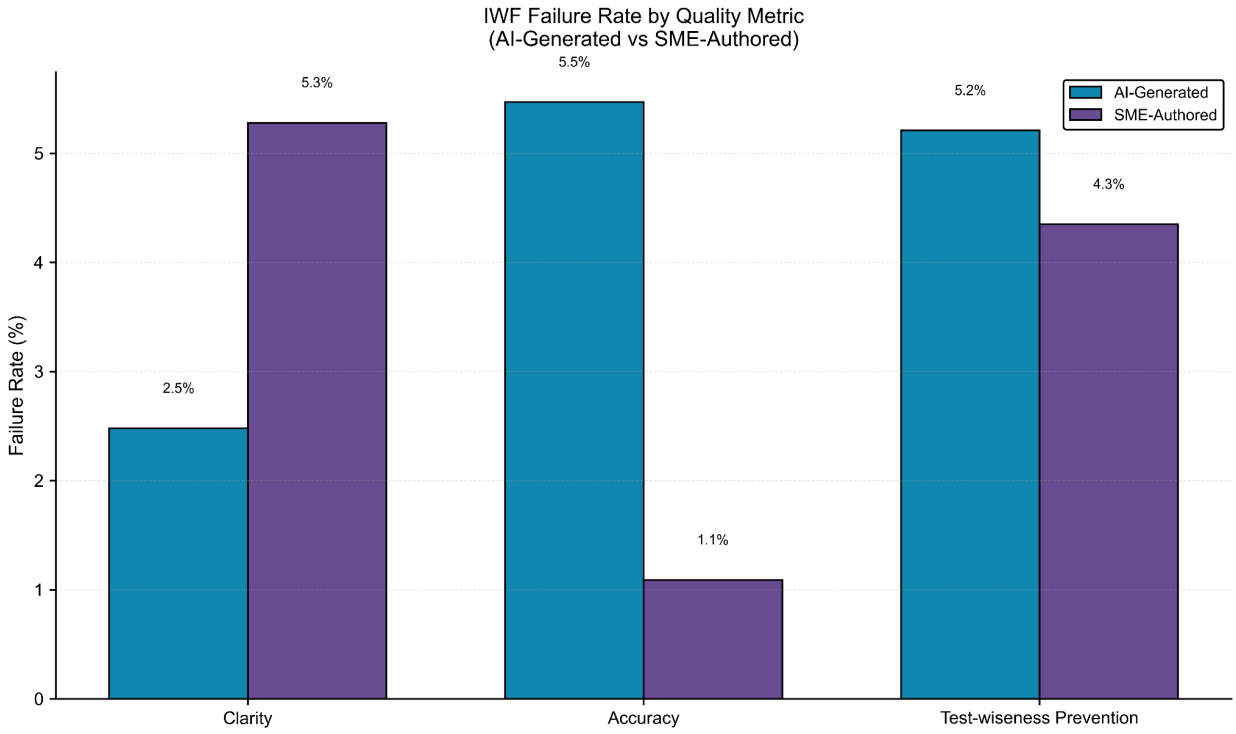

Figure 4: IWF failure rate by quality metric

For the three quality metrics defined above, we measured the failure rate across all three disciplines. We found that the SME-authored questions fail more frequently than AI-generated questions in terms of clarity (5.3% versus 2.5%). However, the AI-generated questions have a higher failure rate for accuracy. Only 1.1% SME-authored questions failed in the context of accuracy. AI-generated questions fail slightly more frequently than SME-authored questions in relation to test-wiseness prevention, meaning that AI-generated questions are slightly more likely to include inadvertent cues that point to the correct answer.

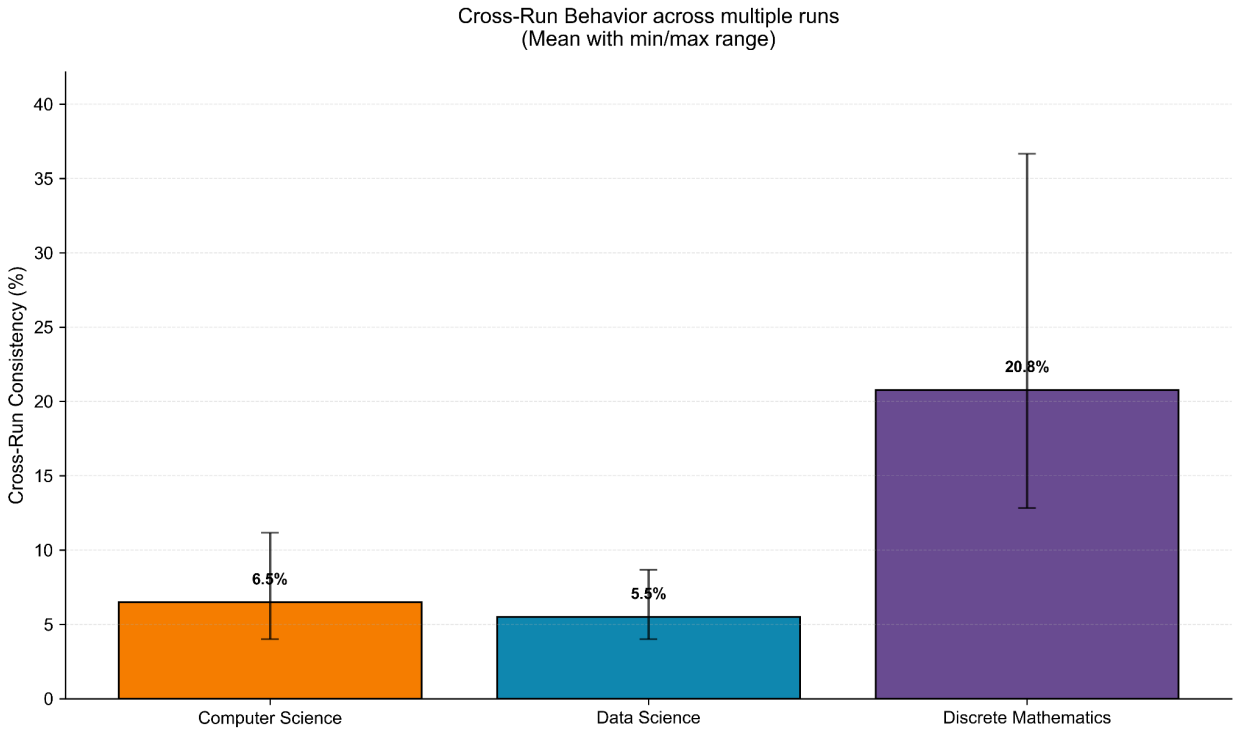

Figure 5: Cross-run behavior across multiple runs

An analysis of the behavior of the tool across multiple runs shows that Discrete Mathematics has the highest rate (20.8) of similarity between runs. Computer science showed about 6.5% similarity and data science about 5.5%. Our initial inference is that the high similarity in discrete math is due to the nature of the subject leading to more procedural questions. This suggests that the nature of the discipline and question types may influence AI generation patterns, warranting further investigation into discipline-specific prompt engineering strategies.

Conclusion and Future Work

In this paper we present an AI tool to generate MCQs and use it to generate questions for computer science, data science, and math higher education content. We also provide a comprehensive evaluation framework to assess the effectiveness of these questions. The analysis consisted of over 2000 questions across the three disciplines and successfully demonstrated that GenAI can reliably produce high-quality assessments when guided by structured prompts that incorporated explicit guardrails and learning science principles.

Our findings carry important implications for educational practice. First, they provide empirical evidence that GenAI can serve as a reliable partner in assessment design when proper guardrails are implemented, addressing instructor concerns about question quality and reliability. Second, the framework we present offers a replicable methodology for evaluating AI-generated assessments across disciplines, enabling institutions to establish quality benchmarks for

AI-assisted content creation. Third, our results suggest that the optimal approach combines AI efficiency with instructor expertise, where GenAI rapidly generates question drafts and subject matter experts provide quality assurance, particularly for technical accuracy.

While our framework currently evaluates quality and relevance, it does not assess actual student performance or learning outcomes, which represents an important direction for future work. We intend to extend the work presented in [6] to incorporate this AI tool to provide students a self-assessment tool intended to help students study. We intend for future research to also extend this framework to other assessment types, examine student performance data to validate the pedagogical effectiveness of AI-generated questions, and explore adaptive generation strategies that adjust difficulty and complexity based on course level and student needs.

In conclusion, this work demonstrates that GenAI, when properly constrained and guided, can generate effective multiple choice questions that meet established quality standards and align with learning objectives. While human oversight remains essential, particularly for ensuring technical accuracy, the combination of AI efficiency and instructor expertise presents a promising path forward for addressing the persistent challenge of assessment creation in higher education. As GenAI capabilities continue to advance, the frameworks and best practices established through studies like this will be crucial for ensuring that AI serves as a genuine enhancement to educational practice rather than a source of additional burden or risk.

Appendix A
AI-generated multiple-choice questions by subject area

The following questions were generated across four independent runs of the same prompt for each of the three subject areas. Each run produced a set of multiple-choice questions on the given topic.

Computer Science - Detecting equal values with branches

Run 1

Q1. What does the equality operator == do in an if statement?

a) Executes the else branch automatically
b) Checks if two values are not equal
c) Compares if two values are equal
d) Assigns a value to a variable

Answer: C

Q2. What is the purpose of the else branch in an if-else statement?

a) Runs code when the if condition is false
b) Compares multiple conditions in sequence
c) Runs code when the if condition is true
d) Checks if two values are equal

Answer: A

Q3. In a multi-branch if-else statement, how many branches can execute?

a) The last branch only
b) Only the first branch that evaluates to True
c) None unless all conditions are True
d) All branches that evaluate to True

Answer: B

Q4. The inequality operator != evaluates to True when two values are _____.

a) Both zero
b) Both non-zero
c) Equal
d) Not equal

Answer: D

Run 2

Q1. What does the equality operator (==) check in an if statement?

a) Whether a value is greater than another
b) Whether a value is less than another
c) Whether two values are equal
d) Whether two values are not equal

Answer: C

Q2. What happens when an if statement's condition evaluates to False?

a) The program crashes
b) The program skips the branch statements
c) The else branch always executes
d) The indented branch statements execute

Answer: B

Q3. In an if-else statement, when does the else branch execute?

a) When the condition uses the inequality operator
b) When the if condition is not written
c) When the if condition is True
d) When the if condition is False

Answer: D

Q4. What keyword is used to add additional conditions after an if statement?

a) elseif
b) elif
c) else
d) then

Answer: B

Run 3

Q1. What does the equality operator (==) check in an if statement?

a) If two values are not equal
b) If a value is false
c) If a value is true
d) If two values are equal

Answer: D

Q2. What happens when an if statement's expression evaluates to False?

a) The program repeats the if statement
b) The statements inside the if branch are skipped
c) The statements inside the if branch execute
d) The program terminates immediately

Answer: B

Q3. In an if-else statement, which branch executes if the condition is False?

a) Neither branch
b) Both if and else branches
c) The else branch

d) The if branch

Answer: C

Q4. What keyword introduces an additional branch between if and else?

a) elif
b) else: if
c) elseif
d) else if

Answer: A

Run 4

Q1. What does the equality operator (==) check in an if statement?

a) If the left and right values are equal
b) If the left and right values are not equal
c) If the left value is assigned to the right value
d) If the left value is greater than the right value

Answer: A

Q2. In an if-else statement, what happens if the condition is false?

a) Both branches execute
b) The else branch executes
c) The program stops running
d) The first branch executes

Answer: B

Q3. What keyword allows checking multiple conditions in sequence in Python?

a) else
b) loop
c) elif
d) switch

Answer: C

Q4. What Boolean value results from the expression: 5 != 5?

a) False

b) True
c) None
d) 5

Answer: A

Data Science - Cleaning data

Q1. What type of dirty data is a numeric value far from the mean called?
a) Duplicate data
b) Missing data
c) Inapplicable data
d) Outlier data

Answer: D

Q2. Which technique removes entire rows with missing values?
a) Pairwise discarding
b) Discarding instances
c) Imputing data
d) Discarding features

Answer: B

Q3. Which pandas method replaces missing values with specified values?
a) df.dropna()
b) df.drop_duplicates()
c) df.replace()
d) df.fillna()

Answer: D

Q4. What is the main disadvantage of pairwise discarding?
a) It introduces bias
b) It duplicates data
c) It removes entire features
d) It causes varying instance counts across analyses

Answer: D

Run 2

Q1. What term describes data values that are much larger or smaller than most values in a feature?

a) Outlier data
b) Missing data
c) Duplicate data
d) Imputed data

Answer: A

Q2. Why should duplicate instances usually be discarded from a dataset?

a) They reduce dataset size unnecessarily
b) They remove important features
c) They cause bias by weighting repeated values too heavily
d) They increase missing data

Answer: C

Q3. When is discarding instances a suitable method to handle dirty data?

a) When dirty instances comprise a small, random percentage of the dataset
b) When dirty instances are correlated with other features
c) When features contain missing values
d) When dirty instances affect more than 60% of features

Answer: A

Q4. Fill in the blank: _____ imputation replaces missing values with the mean of a feature, excluding missing and outlier values.

a) Cold-deck
b) Regression
c) Mean
d) Hot-deck

Answer: C

Run 3

Q1. What type of dirty data is a missing value represented as a blank or NaN?

a) Missing data

b) Outlier
c) Duplicate
d) Invalid data

Answer: A

Q2. Why should outlier data typically be removed or corrected in a dataset?
a) They are harder to identify than duplicates
b) They cause bias and skew analysis results
c) They reduce data storage requirements
d) They decrease the number of instances

Answer: B

Q3. When is discarding an entire feature recommended in data cleaning?
a) When the feature has a few outlier values
b) When the feature has 60% or more missing values
c) When the feature has correlated instances
d) When the feature contains many duplicate values

Answer: B

Q4. How does pairwise discarding handle dirty data during analysis?
a) Replaces missing values with mean values
b) Retains dirty instances for some analyses and discards only when a dirty feature is used
c) Discards all dirty instances regardless of analysis
d) Removes dirty features from the dataset entirely

Answer: B

Run 4

Q1. What term describes data values that are much larger or smaller than most in the same feature?
a) Missing data
b) Dirty instance
c) Outlier data
d) Duplicate data

Answer: C

Q2. Dirty data includes missing, outlier, and _____ data.

a) Normal
b) Duplicate
c) Corrupted
d) Imputed

Answer: B

Q3. When should discarding dirty instances be avoided?

a) When dirty instances correlate with other features
b) When dirty instances are a small dataset percentage
c) When dirty instances are random
d) When duplicates are present

Answer: A

Q4. What is the effect of pairwise discarding on dataset analysis?

a) It discards features with any missing values
b) It removes all duplicates from the dataset
c) It causes varying instance counts per analysis
d) It simplifies data comparison

Answer: C

Discrete mathematics - Laws of propositional logic

Run 1

Q1. If two propositions are logically equivalent, substitution in a compound proposition results in _____.

a) a different truth value
b) an unrelated proposition
c) an invalid expression
d) a logically equivalent proposition

Answer: D

Q2. The law $p \rightarrow q \equiv \neg p \vee q$ is called the _____.

a) Conditional identity

b) Complement law
c) Distributive law
d) De Morgan's law

Answer: A

Q3. Which law justifies $\neg (p \vee q) \equiv \neg p \wedge \neg q$?

a) Absorption law
b) Commutative law
c) De Morgan's law
d) Identity law

Answer: C

Q4. The law that states $p \vee p \equiv p$ is the _____.

a) Idempotent law
b) Associative law
c) Domination law
d) Complement law

Answer: A

Run 2

Q1. The law that allows substituting $\neg p \vee q$ for $p \rightarrow q$ is called the _____.

a) De Morgan's law
b) Conditional identity
c) Distributive law
d) Commutative law

Answer: B

Q2. According to De Morgan's laws, $\neg (p \vee q)$ is equivalent to _____.

a) $\neg p \wedge \neg q$
b) $p \wedge q$
c) $\neg p \rightarrow \neg q$
d) $\neg p \vee \neg q$

Answer: A

Q3. The law that justifies $(p \wedge p) \equiv p$ is the _____.

a) Absorption law
b) Idempotent law
c) Domination law
d) Identity law

Answer: B

Q4. Which law states that $p \vee (q \wedge r) \equiv (p \vee q) \wedge (p \vee r)$?

a) Absorption law
b) Commutative law
c) Associative law
d) Distributive law

Answer: D

Run 3

Q1. What law justifies substituting $p \rightarrow q$ with $\neg p \vee q$?

a) De Morgan's law
b) Conditional identity
c) Double negation law
d) Distributive law

Answer: B

Q2. Using the law of propositional logic, $\neg (p \vee q) \equiv$ _____.

a) $p \vee q$
b) $\neg p \wedge \neg q$
c) $p \wedge q$
d) $\neg p \vee \neg q$

Answer: B

Q3. Which law states that $(p \vee F \equiv p)$?

a) Identity law
b) Complement law
c) Domination law
d) Absorption law

Answer: A

Q4. Which law allows rearranging $\neg p \wedge q$ to $q \wedge \neg p$?

a) Distributive law
b) Idempotent law
c) Commutative law
d) Associative law

Answer: C

Run 4

Q1. The law that justifies $p \rightarrow q \equiv \neg p \vee q$ is the _____.

a) Conditional identity
b) Distributive law
c) Complement law
d) De Morgan's law

Answer: A

Q2. Which law allows substitution of $\neg p \vee q$ for $p \rightarrow q$ in a compound proposition?

a) Commutative law
b) Identity law
c) Conditional identity
d) Absorption law

Answer: C

Q3. The equivalence $\neg (p \vee q) \equiv \neg p \wedge \neg q$ is an example of _____.

a) Complement law
b) Associative law
c) De Morgan's law
d) Distributive law

Answer: C

Q4. Which law states that $p \wedge p \equiv p$?

a) Absorption law
b) Idempotent law
c) Domination law

d) Identity law

Answer: B

Appendix B
Example evaluation output

The following illustrates the evaluation pipeline used to determine classroom-ready status, with one example from each possible outcome category.

Classroom ready - Acceptable quality, aligned to learning objectives

Question: What does the equality operator == do in an if statement?

e) Executes the else branch automatically
f) Checks if two values are not equal
g) Compares if two values are equal
h) Assigns a value to a variable

Answer: C

IWF criteria results: 0 failures (19/19 Pass)

LO alignment result: Yes, this question directly assesses the learning objective of evaluating expressions involving equality operators to produce Boolean outcomes, as it specifically asks about the function of the equality operator in an if statement.

Outcome: Classroom ready.

Refine quality - Unacceptable quality, aligned to learning objectives

Question: In a multi-branch if-else statement, how many branches can execute?

a) The last branch only
b) Only the first branch that evaluates to True
c) None unless all conditions are True
d) All branches that evaluate to True

Answer: B

IWF criteria results: 2 failures - absolute terms, longest option correct

LO alignment result: Yes, the question directly assesses the understanding of multi-branch if-else statements and which branch executes based on conditions, aligning with the learning objectives regarding the use of the elif keyword and determining which branch executes.

Outcome: Not classroom ready. The question requires quality revision before use.

Realign content - Acceptable quality, unaligned to learning objectives

Question: What does an expression in programming always produce?

a) A sequence of instructions
b) A value
c) A variable name
d) A comment

Answer: B

IWF criteria results: 1 failure - absolute terms

LO alignment: No, the question does not directly assess the learning objectives, as it focuses on the general concept of expressions in programming rather than specifically evaluating or writing arithmetic expressions, applying precedence rules, or converting equations to Python statements.

Outcome: Not classroom ready. The question requires a revision to align with the intended learning objectives.

Rewrite - Unacceptable quality, unaligned to learning objectives

Question: In programming, the multiplication symbol is _____, not the × symbol.

a) x
b) ×
c) *
d) ÷

Answer: C

IWF criteria results: 2 failures - negatively worded, convergence cues

LO alignment: No, this question does not directly assess the learning objectives as it focuses on identifying the correct multiplication symbol rather than writing expressions, applying precedence rules, or converting equations to Python statements.

Outcome: Not classroom ready. The question requires both quality revision and realignment with the intended learning objectives.

References

[1] LearnLM Team, “LearnLM: Improving Gemini for learning,” arXiv:2412.16429, 2024. [Online]. Available: https://arxiv.org/abs/2412.16429

[2] K. Satheesh, et al., “AI-powered MCQ generation system: A smart web application for converting study materials into MCQs,” Journal of Computer Science, [Online]. Available: https://computersciencejournal.org/wp-content/uploads/19-JCS1593.pdf

[3] T. M. Haladyna, S. M. Downing, and M. C. Rodriguez, “A review of multiple-choice item-writing guidelines for classroom assessment,” Applied Measurement in Education, vol. 15, no. 3, pp. 309–334, 2002.

[4] J. Breakall, C. Randles, and R. Tasker, “Development and use of a multiple-choice item writing flaws evaluation instrument in the context of general chemistry,” Chemistry Education Research and Practice, vol. 20, no. 2, pp. 369–382, 2019.

[5] S. Moore, E. Costello, H. A. Nguyen, and J. Stamper, “An automatic question usability evaluation toolkit,” arXiv:2405.20529, May 2024.

[6] A. Hui, et al., “Positive student impacts of an unlimited, randomized self-assessment quiz per chapter: Study habits, self-efficacy, and learning outcomes,” zyBooks, 2025. [Online]. Available: https://www.zybooks.com/wp-content/uploads/2025/07/47321-Positive_Student_Impacts_of_an_Unlimited_Randomized_Self-Assessment_Quiz_Per_Chapter__Study_Habits_Self-Efficacy_and_Learning_Outcomes.pdf