

AI-Powered Educational Assessments for Quality Evaluation

Balamurugan M¹, Sravan PP²

¹Associate Professor, Department of MCA, Acharya Institute of Graduate Studies, Bengaluru.

²Department of MCA, Acharya Institute of Graduate Studies, Bengaluru.

DOI: <https://doi.org/10.51583/IJLTEMAS.2026.150900154>

Received: 28 September 2026; Accepted: 03 October 2026; Published: 24 October 2026

ABSTRACT

The growing use of Generative Artificial Intelligence in education has created new opportunities for automating the development of examination questions and assessments. Large Language Models can generate multiple choice, short answer, and descriptive questions from course materials within minutes, reducing the workload of instructors. However, generating valid questions individually does not guarantee that a complete assessment is educationally effective. An AI generated question paper may still have an uneven distribution of difficulty levels, limited coverage of higher order cognitive skills, an imbalance in question types, or multiple questions that assess similar concepts. These issues can reduce the overall quality and diagnostic value of an assessment.

This paper proposes a Quality Evaluation Module that evaluates a complete AI generated question set as a whole before publication. Rather than focusing only on individual questions, the module examines four dimensions: difficulty distribution, coverage across Bloom's Taxonomy levels, balance among question types, and semantic redundancy. Each dimension is evaluated and combined to produce an overall quality score, along with recommendations that help instructors improve the assessment.

The module is designed to integrate with AI powered educational assessment platforms where instructors can upload study materials and use Large Language Models to generate and manage questions. It introduces an additional assessment level evaluation to determine whether the selected questions form a balanced and diverse question paper. The approach is based on research in AI assisted question generation, Bloom's Taxonomy alignment, and assessment quality evaluation.

The results obtained from three sample assessments show that the module can help improve the overall structure and quality of an assessment. After applying its recommendations, the question sets showed a better balance in difficulty, improved coverage of higher cognitive levels, and fewer questions with similar meanings. The overall quality scores also improved. The instructor can look at the test change anything that needs changing and choose if it's good enough to use.

Keywords: Artificial Intelligence, Educational Assessment, Assessment Quality, Large Language Models, Bloom's Taxonomy, Automatic Question Generation.

INTRODUCTION

Educational assessment plays an important role in teaching and learning. Tests, quizzes, and assignments help teachers evaluate students' knowledge, understanding, and ability to apply what they have learned. Creating assessments can be time consuming because teachers must prepare questions, set appropriate difficulty levels, align them with learning objectives, and review the final assessment. This becomes more challenging when assessments are needed for multiple classes or groups.

Generative Artificial Intelligence and Large Language Models (LLMs) have created new possibilities in this area. Modern LLMs can analyse study materials and generate questions in different formats and difficulty

levels much faster than traditional methods. Previous research shows that LLM based question generation can reduce the effort required to prepare quizzes and examinations.

However, generating valid questions individually does not guarantee a well-designed assessment. A generated exam may contain too many easy recall questions and insufficient questions that test application or higher-level thinking. It may also focus mainly on the lower levels of Bloom's Taxonomy, such as Remember and Understand, while giving limited attention to Apply, Analyse, Evaluate, and Create. Another important issue is redundancy. When many questions are generated from the same study material, several questions may test the same concept in slightly different ways. This reduces the diversity and effectiveness of the assessment.

These challenges show the need to evaluate the complete assessment rather than only individual questions. To address this problem, this paper introduces a Quality Evaluation Module for AI generated assessments. The module evaluates a question set using four aspects: difficulty distribution, Bloom's Taxonomy coverage, balance among question types, and semantic redundancy. Each aspect is given a score, and the scores are combined to produce an overall quality score. The module also provides suggestions to help instructors identify areas that may need improvement.

LITERATURE REVIEW

Xia, Weng, Ouyang, Lin and Chiu (2024) carried out a scoping review to see how Generative AI changes assessment practice in education. They gathered studies that look at the views of students, teachers and institutions. This review shows chances such as automated assessment help, personalised learning and quick feedback. It also points out problems like how to use Generative AI whether teachers are ready and how schools need new policies for assessment redesign. A limitation of this study is it does not offer a way to judge the quality of an AI-generated test. The work focuses on this gap. It gives a workable method to evaluate the quality of AI-generated exams.

Ling and Afzaal (2024) examined pre-trained LLMs for generating question–answer pairs in higher education, comparing pipeline, joint, and multi-task generation approaches, and evaluating outputs through a combination of automated metrics, teacher judgement, and real classroom use. Their results show that LLM-generated question–answer pairs can meaningfully support learning while reducing the effort required for assessment preparation, making this approach relevant to AI-powered educational assessment systems where questions are automatically generated from learning materials. A limitation of this study is generating and evaluating individual question–answer pairs; it does not ask whether a collection of such pairs forms a balanced examination as a whole.

Elkins, Kochmar, Cheung and Serban (2024) studied how teachers can use LLMs together with Bloom's Taxonomy to build educational quizzes, generating questions tied to specific Bloom's-aligned learning goals and evaluating how teachers used the results in practice. They report that teachers preferred the AI-assisted workflow and found no loss of quiz quality compared with fully handwritten versions.

A Limitation of this study is to focus stays on generating and using questions tied to individual learning goals, not on measuring the cognitive-level distribution of an entire finished assessment. This paper adopts Bloom's Taxonomy as one of four evaluation dimensions rather than as the sole generation criterion.

Research Gap

Existing studies have explored different aspects of AI-assisted assessment, including the use of Generative AI in higher education, automatic question generation, Bloom's Taxonomy-based quiz creation, topic diversity, and the quality of AI-generated exams. However, these studies mainly focus on individual questions, generation methods, or broad assessment practices, rather than providing a unified method for evaluating the quality of a complete AI-generated examination. There is a need for a practical and reusable evaluation mechanism that can assess an entire question paper using multiple criteria at the same time.

Therefore, this study addresses this gap by proposing

an integrated assessment-quality evaluation approach based on difficulty, Bloom's Taxonomy coverage, question-type balance, and question redundancy, enabling instructors to identify quality issues and improve AI-generated examinations before they are used with students.

METHODOLOGY

Dataset

The dataset consists of educational material drawn from selected academic subjects, including course notes, textbook excerpts, lecture material, and study documents. This material is processed to extract clean textual content, which is then used to generate a pool of candidate questions. Rather than treating each question as a separate record, the dataset is organized into full assessments because the main focus of this study is the assessment as a whole, not each item.

The following attributes are attached to each question and are fed directly into the evaluation module:

- Question text
- Question type (MCQ, short answer, long answer)
- Difficulty level (Easy, Medium, Hard)
- Bloom's Taxonomy level (Remember, Understand, Apply, Analyse, Evaluate, Create)
- Marks allotted
- Reference to the source material it was generated from

Data pre-processing

Before evaluation, the raw material and generated questions pass through a pre-processing pipeline consisting of the following stages:

- Text Extraction - extracting raw text from uploaded PDF study materials.
- Text Cleaning - removing extraneous whitespace, repeated line breaks, and formatting artefacts.
- Question Standardization - converting generated questions into a consistent structured record (text, type, difficulty, Bloom's level, marks, source reference).
- Duplicate-Detection Preparation - normalizing and embedding question text so that it is ready for semantic similarity comparison.
- Metadata Validation - checking that every question record carries a valid difficulty level, question type, and Bloom's Taxonomy tag before it is admitted to the evaluation stage.
- Assessment Formation - grouping validated questions into the specific sets an instructor has assembled, since quality is evaluated at the level of the assessment, not the individual question.

Proposed Method

The core contribution of this paper is a module that evaluates a complete AI-generated assessment immediately before publication, rather than validating questions one at a time. The module is organized into

four analytical components, each targeting a distinct quality concern.

Difficulty Distribution Analysis

The module counts how many Easy, Medium and Hard questions are in the assessment. The module also measures what percentage of each type appears. The module then compares that observed distribution to a target distribution that the instructor sets or that the platform provides. Table 4.1.1 shows an example that looks like this. When the module finds that the difference between the generated percentages and the target percentages is larger than a set limit the module flags that dimension.

Difficulty	Generated (%)	Target (%)
Easy	50	30
Medium	35	50
Hard	15	20

Table 4.1.1: Example comparison of generated vs. target difficulty distribution.

When the module finds that the difference between the generated percentages and the target percentages is larger than a set limit the module flags that dimension. Recommends a fix, such, as adding more medium questions or swapping out too many Easy questions.

Bloom's Taxonomy Coverage Analysis

Each question has a tag that shows which level of Bloom's Taxonomy it belongs to. Remember, Understand, Apply, Analyse, Evaluate or Create. And the module looks at the test to see if there are enough questions from each of these thinking levels. If some levels are not there or if some are too common (Remember and Understand, which often take up too much space compared to Apply, Analyse, Evaluate and Create), the system points this out clearly to the teacher

Question-Type Balance Analysis

The module tallies the distribution of question types MCQ, short answer, long answer and checks it against the mixture an instructor has specified for the assessment, as illustrated in Table 4.1.3.

Question Type	Number of Questions
MCQ	15
Short Answer	5
Long Answer	3

Table 4.1.3: Example question-type distribution for a generated assessment.

Semantic Redundancy Analysis

Semantic similarity techniques are used to compare every pair of questions in the assessment. Pairs whose similarity exceeds a set threshold, indicating that they test substantially the same concept, are flagged as potentially redundant. For each flagged pair, the instructor is offered the option to remove one question, replace a question, request regeneration of a question, or retain both if the overlap is judged acceptable.

Overall Quality Score

The outputs of the four components are normalised and combined into a single, interpretable Assessment Quality Score:

Assessment Quality Score = f (Difficulty Balance Score, Cognitive Coverage Score, Question-Type Balance Score, Diversity Score)

Each of the four component scores is adjusted to a 0–100 scale before being combined. The institution can combine them using an average, with weights it chooses. The result is a number that reflects the overall quality of the assessment composition. Alongside this number the four individual sub-scores are also shown, as seen in Table 4.2.

Quality Dimension	Score (/100)
Difficulty Balance	85
Bloom's Coverage	80
Question-Type Balance	90
Semantic Diversity	75
Overall Score	82.5

Table 4.2: Example composition of an overall Assessment Quality Score.

Recommendation Generation

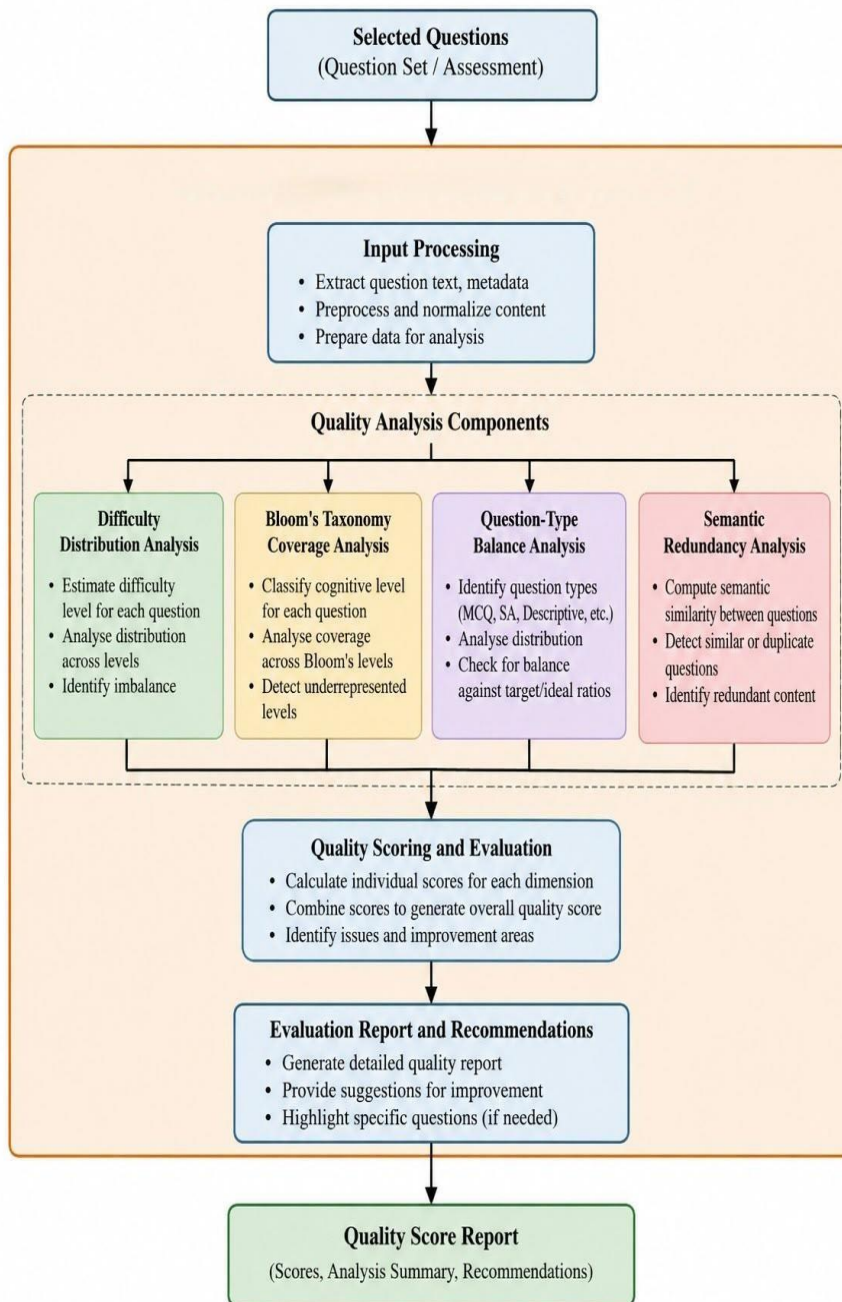
Based on the four analyses, the module generates specific, actionable recommendations rather than a bare numeric score. Typical examples include:

- Add more medium-difficulty questions.
- Increase questions at the Apply and Analyse Bloom's levels.
- Reduce the number of near-duplicate MCQs.
- Replace semantically similar question pairs.
- Include more descriptive (short/long answer) questions.

These recommendations are advisory: the module never edits the assessment automatically. The instructor reviews the flagged issues and retains complete control over which changes, if any, to make.

System Architecture

The proposed Quality Evaluation Module receives the complete set of questions selected by the instructor for an assessment. Each question includes the question text, question type, difficulty level, Bloom's Taxonomy level, and marks. The module processes these questions and evaluates the overall assessment rather than checking each question separately. The module performs four main analyses: difficulty distribution, Bloom's Taxonomy coverage, question-type balance, and semantic redundancy. It checks whether the assessment has a suitable mix of difficulty levels, covers different cognitive levels, maintains a balanced distribution of question types, and avoids questions that are too similar.



The results of these analyses are combined to generate an Assessment Quality Score along with specific recommendations for improvement. The instructor can review the quality report, make necessary changes to the assessment, and evaluate it again if required. The module only provides recommendations and does not automatically modify the assessment. Once the instructor is satisfied with the final question set, the assessment can be published for students.

Figure 4.4: System architecture

RESULTS AND DISCUSSION

Difficulty Distribution

The module was evaluated on whether applying its recommendations moves a generated assessment closer to its intended difficulty distribution. Table 5.1 and Figure 5.1 summarise an assessment that initially over-produced easy questions are brought in line with a 30/50/20 (Easy/Medium/Hard) target after instructor's act on the module's recommendations.

Difficulty Level	Before Evaluation (%)	After Improvement (%)	Target (%)
Easy	50	30	30
Medium	30	50	50
Hard	20	20	20

Table 5.1: Difficulty distribution before and after applying module recommendations.

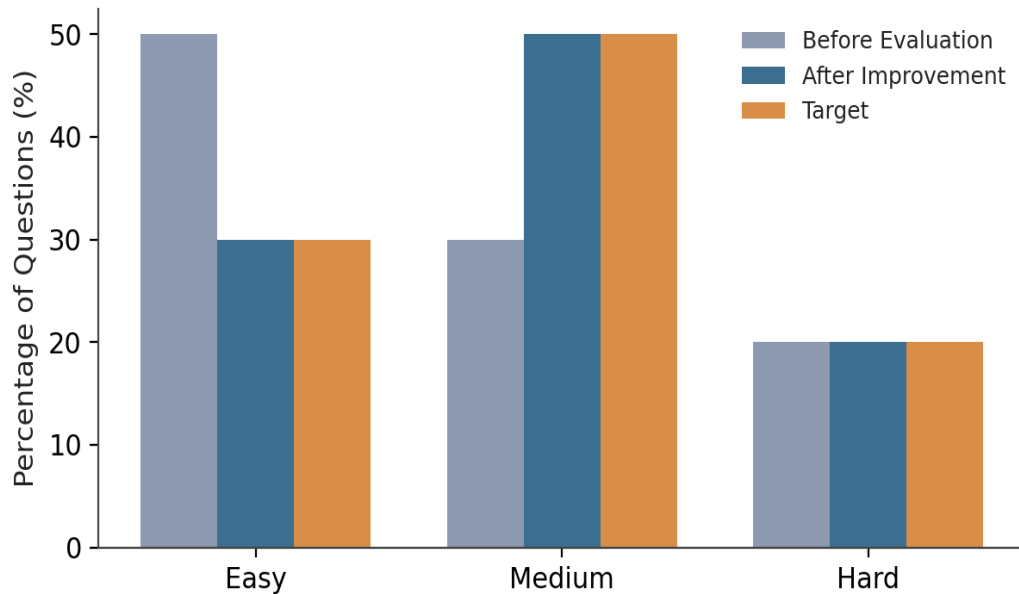


Figure 5.1: Difficulty Distribution – Before vs After vs Target

Bloom's Taxonomy Coverage

The module's second evaluation checks how well an assessment covers the six Bloom's Taxonomy levels. In the Table 5.2, Figure 5.2, the initial assessment is dominated by Remember and Understand items, with almost no representation of Analyse, Evaluate, or Create. After applying the module's recommendations, coverage shifts noticeably toward the higher-order levels, indicating an assessment that tests deeper understanding rather than recall alone.

Table 5.2: Bloom's Taxonomy coverage before and after applying module recommendations

Bloom's Level	Before	After
Remember	8	5
Understand	7	6
Apply	3	6
Analyse	1	3
Evaluate	0	1
Create	0	1

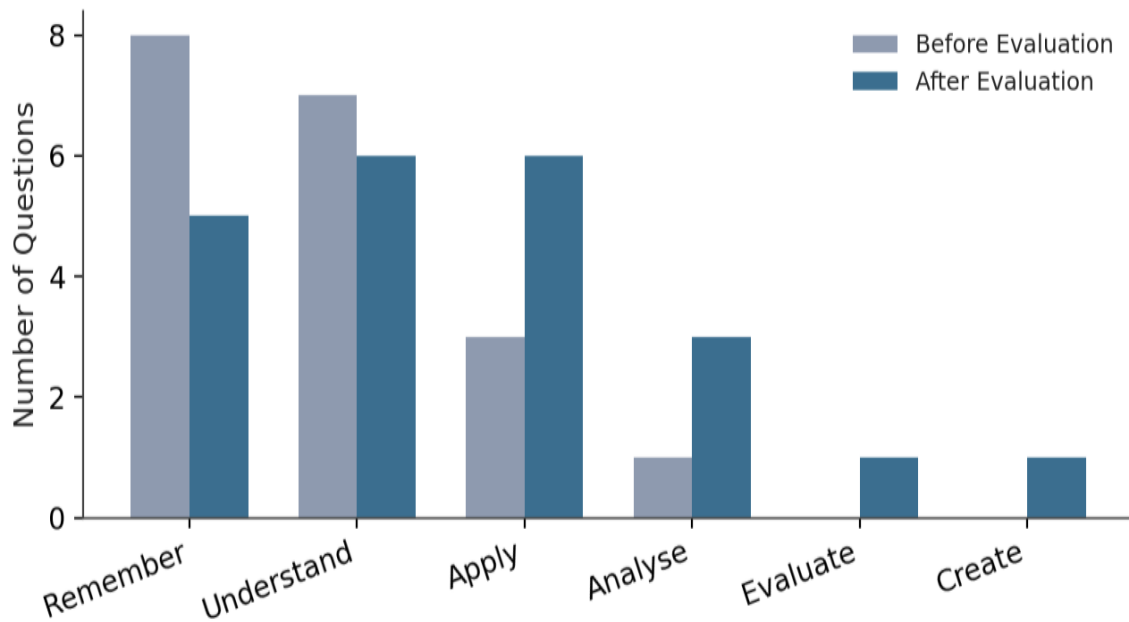


Figure 5.2: Bloom's Taxonomy Coverage – Before vs After

Semantic Redundancy

Semantic similarity analysis was used to identify potentially redundant question pairs within an assessment. In Table 5.3, Figure 5.3, the number of flagged redundant questions falls from 6 to 2 after the instructor acts on the module's redundancy recommendations, indicating a meaningful improvement in topical diversity.

Assessment Version	Redundant Questions
Before Evaluation	6
After Evaluation	2

Table 5.3: Reduction in flagged redundant questions.

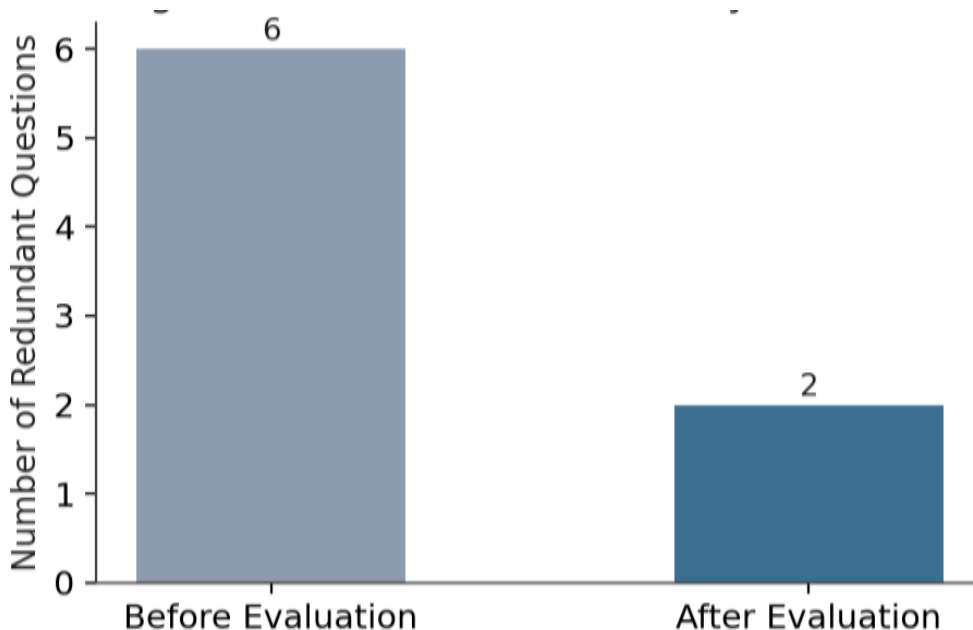


Figure 5.3: Semantic Redundancy Reduction

Overall Quality Comparison

Finally, the four component scores were combined into an overall Assessment Quality Score for three assessments, both before and after applying the module's recommendations (Table 5.4, Figure 5.4). Across all three, the overall score rises from the high-60s / low-70s range into the low-to-mid 80s, consistent with the improvements observed in the individual dimensions above.

Assessment	Initial Quality Score	Final Quality Score
Assessment 1	68	85
Assessment 2	72	88
Assessment 3	65	82

Table 5.4: Overall quality score comparison across three assessments

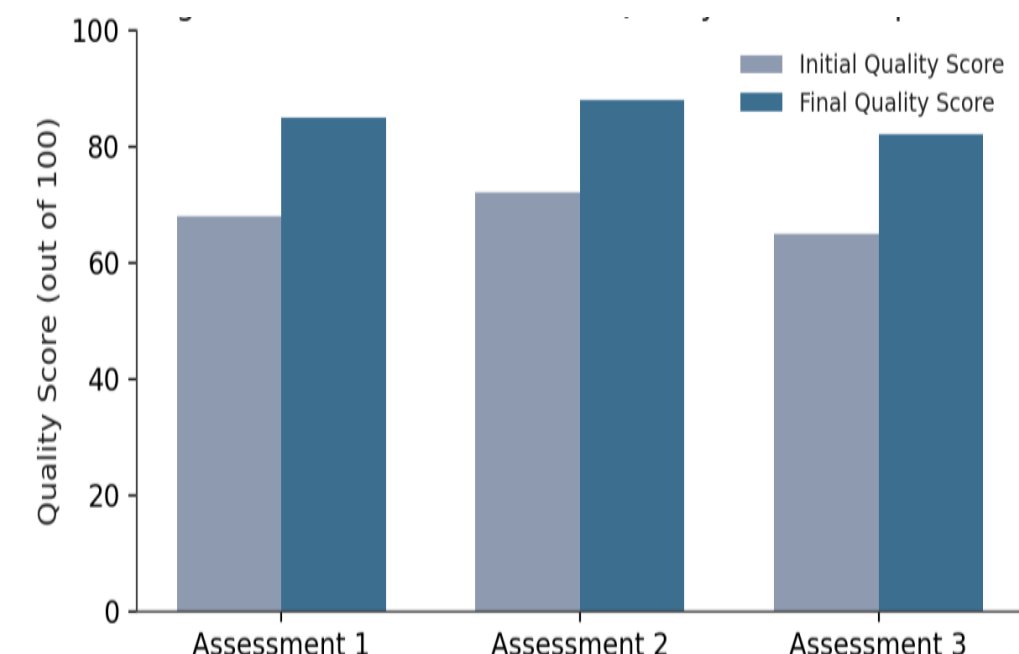


Figure 5.4: Overall Assessment Quality Score Comparison

Overall, these results show that the proposed module can identify important problems in an AI-generated assessment, such as an unbalanced difficulty level, limited coverage of higher levels of Bloom's Taxonomy, and repeated or similar questions. The module can also provide useful suggestions to address these problems and improve the overall quality score of the assessment. The results indicate that evaluating the complete assessment provides a better view of its overall structure and quality. The quality score helps instructors quickly understand the strengths and weaknesses of the assessment. This can support instructors in making informed changes before publishing the assessment to students.

CONCLUSION

Generative Artificial Intelligence carries real potential to ease the burden of educational assessment creation, cutting the time instructors spend writing questions and supporting the rapid development of exams and quizzes. However, this study's central argument is that generating individually valid questions is not the same as producing a balanced, diverse, and pedagogically appropriate assessment as a whole.

This paper has proposed a Quality Evaluation approach that closes that gap by examining a complete AI-generated assessment along four dimensions difficulty distribution, Bloom's Taxonomy coverage, question-type balance, and semantic redundancy before it is published. The principal contribution is the introduction of an integrated, pre-publication, assessment-level quality-evaluation mechanism: rather than adding another question-generation technique, this work adds a checkpoint that evaluates the composition of an assessment already assembled from generated questions, returning a single quality score together with concrete recommendations.

The proposed module is designed to be integrated directly into AI-powered assessment platforms, where questions are generated from uploaded educational materials and organised into assessments, providing instructors with an additional layer of quality assurance before the assessment is presented to students. Future work may extend this approach through psychometric analysis, mapping to curriculum outcomes, fairness analysis across student subgroups, instructor-specific assessment preferences, and adaptive, ongoing assessment evaluation as question banks grow over time.

REFERENCES

1. Chen, Y., Wu, C., Yan, S., Liu, P., & Xiao, Y. (2024), "Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics", Vol.1, pp. 3138–3167. <https://doi.org/10.18653/v1/2024.acllong.173>
2. Deroy, A., Maity, S., & Sarkar, S. (2025), "Proceedings of Large Foundation Models for Educational Assessment", (Vol. 264, pp. 3–32).
3. Elkins, S., Kochmar, E., Cheung, J. C. K., & Serban, I. (2024), "Proceedings of the AAAI Conference on Artificial Intelligence", Vol. 38(21), pp. 23084–23091.
4. Fu, W., Wei, B., Hu, J., Cai, Z., & Liu, J. (2024), "QGEval: Benchmarking multi-dimensional evaluation for question generation", Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 11783–11803). <https://doi.org/10.18653/v1/2024.emnlp-main.658>
5. Gorgun, G., & Bulut, O. (2024), "Current evaluation methods are a bottleneck in automatic question generation", Proceedings of the 2024 AAAI Conference on Artificial Intelligence (Vol. 257, pp. 3–8).
6. Isley, C., Gilbert, J., Kassos, E., Kocher, M., Nie, A., Brunskill, E., Domingue, B., Hofman, J., Legewie, J., Svoronos, T., Tuminelli, C., & Goel, S. (2026), "Assessing the quality of AI-generated exams: A large-scale field study". Proceedings of the AAAI Conference on Artificial Intelligence, 40(45), 38626–38634. <https://doi.org/10.1609/aaai.v40i45.41205>
7. Jung, J. Y., Tyack, L., & von Davier, M. (2025), "Towards the implementation of automated scoring in international large-scale assessments: Scalability and quality control" Computers and Education: Artificial Intelligence, 8, Article 100375. <https://doi.org/10.1016/j.caeai.2025.100375>
8. Ling, J., & Afzaal, M. (2024). "Automatic question-answer pairs generation using pre-trained large language models in higher education" Computers and Education: Artificial Intelligence, 6, Article 100252. <https://doi.org/10.1016/j.caeai.2024.100252>
9. Nikolovski, V., Trajanov, D., & Chorbev, I. (2025). "Advancing AI in higher education: A comparative study of large language model-based agents for exam question generation, improvement, and evaluation". Algorithms, 18(3), Article 144. <https://doi.org/10.3390/a18030144>
10. Sairaj, R. T., & Balasundaram, S. R. (2025), "Ensemble learning with RAG model to reduce redundant question topics in auto-generated exam questions", Discover Computing, 28, 161. <https://doi.org/10.1007/s10791-025-09683-2>
11. Scaria, N., Chenna, S. D., & Subramani, D. (2024), "How good are modern LLMs in generating relevant and high-quality questions at different Bloom's skill levels", Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024) (pp. 1–10).
12. Villegas-Ch, W., Gutierrez, R., García-Ortiz, J., & Guevara, V. (2025) "Explainable educational assistant integrated in Moodle: Automated semantic assessment and adaptive tutoring based on NLP and XAI", Discover Artificial Intelligence, 5, Article 191. <https://doi.org/10.1007/s44163-025-00438-y>
13. Witsken, G., Crk, I., & Gultepe, E. (2025). "LLMs in the classroom: Outcomes and perceptions of questions written with the aid of AI". Proceedings of the AAAI Conference on Artificial Intelligence,

- 39(26), 27698–27705. <https://doi.org/10.1609/aaai.v39i26.34984>
14. Xia, Q., Weng, X., Ouyang, F., Lin, T. J., & Chiu, T. K. F. (2024). “A scoping review on how generative artificial intelligence transforms assessment in higher education”, *International Journal of Educational Technology in Higher Education*, 21, Article 40. <https://doi.org/10.1186/s41239-024-00468-z>
15. Yaacoub, A., Da-Rugna, J., & Assaghir, Z. (2025). “Assessing AI-generated questions’ alignment with cognitive frameworks in educational assessment”, *International Journal of Computer Theory and Engineering*, 17(3), 14–125 <https://doi.org/10.7763/IJCTE.2025.V17.1374>