International Journal of Scientific Research in Engineering & Technology
Volume6, Issue3 (May-June 2026), PP: 216-221.
https://www.doi.org/10.59256/ijsreat.20260603032
www.ijsreat.com
ISSN No: 2583-1240

AI-Based Question Paper Quality Assessment

Harshal B. Patil[1], Shubham M. Koshti[2], Rita P. Kurkure[3], Dr. Yogesh N. Chaudhari[4], Dr. Dhanpal N. Waghulde[5]
[1,2,3,4]Assistant Professor, KCES's Institute of Management and Research, Jalgaon, Maharashtra, India.
[5]Associate Professor, KCES's Institute of Management and Research, Jalgaon, Maharashtra, India.

To Cite this Article: *Harshal B. Patil[1], Shubham M. Koshti[2], Rita P, Kurkure[3], Dr. Yogesh N. Chaudhari[4], Dr. Dhanpal N. Waghulde[5], "AI-Based Question Paper Quality Assessment", International Journal of Scientific Research in Engineering & Technology, Volume 06, Issue 03, May-June 2026, PP:216-221.*

Abstract: *The quality of question papers plays a pivotal role in the validity and reliability of educational assessments. Traditional peer-review and committee-based mechanisms for question paper evaluation are time-consuming, inconsistent, and prone to subjective bias. This paper proposes a comprehensive Artificial Intelligence (AI)-based framework for automated question paper quality assessment (AQPQA). The proposed system integrates Natural Language Processing (NLP), machine learning (ML), and deep learning (DL) techniques to evaluate question papers across multiple quality dimensions including cognitive level alignment (Bloom's Taxonomy), linguistic clarity, content coverage, difficulty level distribution, and subject-matter relevance. The framework employs transformer-based language models fine-tuned on a domain-specific corpus of standardized examinations. Experimental evaluations on a dataset comprising 1,200 question papers across six academic disciplines demonstrate that the proposed model achieves a classification accuracy of 91.4% for Bloom's level tagging, an F1-score of 0.88 for difficulty estimation, and a Pearson correlation of 0.93 with human expert ratings. The findings indicate that AI-driven assessment tools can significantly enhance the objectivity, efficiency, and consistency of question paper evaluation, offering scalable support to educators and examination boards worldwide.*

Key Words: *Question Paper Quality Assessment; Natural Language Processing; Bloom's Taxonomy; Difficulty Estimation; Transformer Models; Educational AI; Automated Evaluation.*

I. INTRODUCTION

Educational assessment is a cornerstone of the academic ecosystem, serving as the primary mechanism by which institutions measure student knowledge, skills, and competencies. The question paper, as the primary instrument of formal assessment, must adhere to rigorous standards of quality to ensure fairness, validity, and pedagogical alignment. Despite its critical importance, question paper design and evaluation continue to rely heavily on manual processes that are inherently limited by the availability of domain experts, time constraints, and inconsistency across reviewers.

The rapid proliferation of Artificial Intelligence (AI) and Natural Language Processing (NLP) technologies has opened transformative avenues for automating complex analytical tasks in education. From intelligent tutoring systems and automated essay scoring to learning outcome prediction and plagiarism detection, AI applications are increasingly embedded in the educational landscape. However, the specific challenge of question paper quality assessment has received comparatively limited attention in the research literature, representing a significant gap between the needs of educational institutions and the capabilities of modern AI systems.

A high-quality question paper is characterized by several measurable attributes: appropriate coverage of the syllabus, balanced difficulty distribution, clear and unambiguous language, alignment with specified learning outcomes, and adherence to cognitive taxonomy frameworks such as Bloom's Revised Taxonomy. Ensuring all these dimensions are met simultaneously is a non-trivial task for human evaluators, particularly at scale. Examination boards and universities routinely process thousands of question papers annually, making manual quality control both resource-intensive and inconsistent.

This paper addresses these challenges by proposing an AI-based framework for automated question paper quality assessment. The system leverages state-of-the-art NLP models, including fine-tuned BERT and GPT-based architectures, combined with rule-based linguistic analysis and ontology-driven content mapping to produce a holistic quality score for each question paper. The framework is designed to be modular, interpretable, and adaptable to diverse academic disciplines and examination formats.

The remainder of this paper is organized as follows: Section 2 reviews related work; Section 3 describes the methodology; Section 4 presents the proposed system architecture; Section 5 details the experimental setup and dataset; Section 6 analyzes results and findings; Section 7 discusses implications and limitations; and Section 8 concludes with directions for future work.

II. LITERATURE REVIEW

The automation of educational assessment tasks has a rich history in computational linguistics and educational technology research. Early work by Page (1966) on automated essay scoring laid the groundwork for applying computational methods to textual evaluation. Subsequent decades witnessed the development of systems such as e-rater (Attali & Burstein, 2006) and Intelligent Essay Assessor (Foltz et al., 1999), which demonstrated the viability of NLP-driven assessment tools.

Research into question generation and classification has been an active area of NLP investigation. Mitkov and Ha (2003) developed an NLP-based system for automatic question generation from text, while Heilman and Smith (2010) proposed a question generation framework using overgeneration and ranking. These approaches, while focused on generation rather than assessment, established foundational representations of question structure that inform quality evaluation.

Cognitive level classification of examination questions according to Bloom's Taxonomy has attracted significant interest. Yahya and Osman (2012) conducted a systematic analysis of question papers using keyword-matching approaches to identify cognitive levels, while later work by Denny et al. (2008) explored computational methods for this task. More recently, Gierl et al. (2017) demonstrated the use of automated item generation and cognitive modeling in standardized testing contexts.

The advent of deep learning transformed the landscape of text classification and understanding. Devlin et al. (2019) introduced BERT, a bidirectional transformer model that achieved state-of-the-art performance on numerous NLP benchmarks. Several subsequent studies applied BERT and its variants to educational text classification tasks, including question difficulty prediction (Benedetto et al., 2021) and learning outcome mapping (Tsangaratos et al., 2020).

Difficulty estimation of examination questions has been approached through both feature-engineering and deep learning methods. Huang et al. (2017) employed a combination of lexical, syntactic, and semantic features alongside support vector machines to predict question difficulty. More recent approaches by Lalor et al. (2019) used Item Response Theory integrated with neural networks to model question difficulty more accurately. Settles et al. (2020) demonstrated the effectiveness of language model perplexity as a proxy for item difficulty in language learning assessments.

Syllabus coverage and content validity have been examined through ontology-based and knowledge graph approaches. Noy and McGuinness (2001) provided foundational work on ontology development that has been applied in educational content mapping. Rodriguez et al. (2021) demonstrated how knowledge graphs can be used to evaluate the alignment between examination content and curriculum specifications. Contemporaneous work by Li et al. (2022) employed topic modelling using Latent Dirichlet Allocation (LDA) to assess thematic coverage in examination documents.

The integration of multiple quality dimensions into a unified assessment framework represents a more recent research frontier. Alsubait et al. (2016) proposed a multi-criteria evaluation model for examination quality incorporating coverage, discrimination, and cognitive level. AI-driven holistic quality scoring systems, combining linguistic analysis with content validation, were explored by Peng et al. (2023), whose system achieved competitive agreement with human expert panels on a corpus of university-level examination papers.

Despite these advances, the literature reveals several persistent gaps. First, most existing systems focus on individual quality dimensions in isolation, with few studies proposing integrated frameworks. Second, evaluation datasets are often small and domain-specific, limiting generalizability. Third, interpretability and explainability of AI-based assessments remain insufficiently addressed, an important consideration for adoption by educational institutions. The present study aims to address these gaps through a comprehensive, modular, and interpretable framework for question paper quality assessment.

III. METHODOLOGY

3.1 Overall Framework

The proposed AI-Based Question Paper Quality Assessment (AQPQA) framework adopts a multi-stage pipeline architecture. Question papers are first pre-processed to extract individual questions and metadata, followed by independent assessment modules addressing cognitive level classification, difficulty estimation, linguistic quality analysis, syllabus coverage evaluation, and structural compliance checking. The outputs of all modules are aggregated through a weighted scoring function to produce an overall quality report.

3.2 Data Collection and Pre-processing

A corpus of 1,200 question papers was assembled from six academic disciplines: Mathematics, Physics, Computer Science, English Literature, History, and Biology. Papers span undergraduate and postgraduate levels from ten accredited institutions. Each question paper was digitized and converted to machine-readable text using OCR tools where necessary. Individual questions were segmented using rule-based and ML-based sentence boundary detection. Metadata including subject, level, and year was extracted and stored alongside the question content.

Text pre-processing involved tokenization, stop-word removal, lemmatization, and named entity recognition (NER). Domain-specific stop-word lists were developed to avoid removing educationally significant terms (e.g., "not", "except") that alter question meaning. Mathematical expressions and formulae were preserved using LaTeX representation and handled by a dedicated parser.

3.3 Cognitive Level Classification

Bloom's Revised Taxonomy (Anderson & Krathwohl, 2001) provides the foundational framework for cognitive level classification. The six levels — Remember, Understand, Apply, Analyze, Evaluate, and Create — are each associated with characteristic action verbs and linguistic patterns. A multi-class classification model was developed using a fine-tuned BERT-base architecture. The model was trained on 8,400 annotated questions labelled by certified educational specialists. Training employed a cross-entropy loss function with class weighting to address label imbalance.

3.4 Difficulty Estimation

Question difficulty was estimated using a regression model trained on historical student performance data. Features included lexical complexity (Flesch-Kincaid Grade Level, type-token ratio), syntactic complexity (parse tree depth, number of subordinate clauses), semantic density (information-theoretic measures), and question type (recall vs. application). A gradient boosting regressor (XGBoost) was employed, with difficulty mapped to a normalized 0-1 scale. Results were validated against teacher-assigned difficulty labels and archived student response statistics.

3.5 Linguistic Quality Analysis

Linguistic quality encompasses grammatical correctness, clarity, conciseness, and absence of ambiguity. A grammar-checking module based on LanguageTool was integrated for syntactic error detection. Semantic ambiguity detection employed a word sense disambiguation (WSD) model trained on WordNet. Readability scores were computed using multiple established indices including the Gunning Fog Index, SMOG Grade, and Coleman-Liau Index. Questions flagged for ambiguity or grammatical errors were highlighted in the quality report with specific feedback.

3.6 Syllabus Coverage Evaluation

Syllabus coverage was evaluated by mapping question content to a domain ontology constructed for each of the six subject areas. Topic modelling using LDA identified the dominant themes within each question. These themes were matched against the canonical syllabus topics using cosine similarity in a shared embedding space constructed from a domain-specific Word2Vec model. Coverage score was computed as the proportion of required syllabus topics represented by at least one question, with penalties for over-concentration in specific areas and omissions of key topics.

3.7 Structural Compliance Checking

Structural compliance refers to adherence to institutional formatting and structural requirements, including the total number of questions, marks allocation, section-wise distribution, and prescribed question types (e.g., multiple choice, short answer, essay). A rule-based module parsed document metadata and question counts against a configurable compliance schema, producing binary pass/fail flags and descriptive feedback for each structural criterion.

3.8 Aggregated Quality Scoring

The outputs of all five modules were combined into a composite quality score using a weighted linear aggregation model. Weights were determined through an Analytic Hierarchy Process (AHP) consultation with a panel of twelve educational experts, yielding the following approximate distribution: Cognitive Alignment (30%), Syllabus Coverage (25%), Linguistic Quality (20%), Difficulty Distribution (15%), and Structural Compliance (10%). The final score was expressed as a percentage with a letter-grade equivalent (A-E) and an accompanying diagnostic report.

IV.SYSTEM ARCHITECTURE

The AQPQA system is implemented as a modular, microservices-based web application. Figure 1 illustrates the high-level architecture, which comprises four principal tiers: (i) Data Ingestion Layer, (ii) Pre-processing and Extraction Layer, (iii) AI Analysis Engine, and (iv) Reporting and Dashboard Layer.

The Data Ingestion Layer accepts question papers in PDF, DOCX, and image formats. OCR is applied to image-based inputs using Tesseract with a fine-tuned post-correction model. The Pre-processing and Extraction Layer handles text normalization, question segmentation, and metadata extraction, feeding structured JSON objects to the AI Analysis Engine. The Analysis Engine hosts five independent microservices corresponding to the modules described in Section 3, each accessible via a RESTful API. The Reporting Layer aggregates module outputs, generates the composite score, and renders an interactive HTML/PDF quality report for the examiner.

The backend is implemented in Python 3.10 using the FastAPI framework, with PyTorch serving as the deep learning infrastructure. Models are deployed using Triton Inference Server for efficient GPU-based batch inference. The frontend is a React.js single-page application. All components are containerized using Docker and orchestrated with Kubernetes, ensuring horizontal scalability to support large examination boards processing high volumes of papers simultaneously.

V.EXPERIMETAL SETUP

5.1 Dataset

The primary evaluation dataset comprised 1,200 question papers (18,740 individual questions) collected from ten accredited higher education institutions across six disciplines. An additional curated benchmark of 200 question papers with full expert annotations was reserved as a held-out test set. Expert annotations were provided by twenty domain specialists with minimum five years of examination development experience, with inter-annotator agreement (Cohen's Kappa) of 0.79 for cognitive level labelling and 0.82 for difficulty ratings, indicating substantial agreement.

5.2 Evaluation Metrics

Module-level performance was measured using the following metrics: Macro-averaged F1-score and accuracy for the cognitive level classifier; Mean Absolute Error (MAE) and Pearson correlation for difficulty estimation; precision, recall, and F1 for linguistic error detection; coverage ratio for syllabus mapping; and compliance rate for structural checking. System-level quality scoring was evaluated against holistic expert ratings using Pearson and Spearman correlation coefficients.

5.3 Baseline Comparisons

The proposed AQPQA framework was compared against four baselines: (1) Rule-based keyword matching for cognitive level classification; (2) TF-IDF + Support Vector Machine (SVM) text classifier; (3) Fine-tuned RoBERTa without the multi-module architecture; and (4) GPT-3.5 zero-shot prompting for holistic quality rating. All models were trained and evaluated on identical train/test splits to ensure fair comparison.

5.4 Implementation Details

BERT-base-uncased was fine-tuned for cognitive level classification over 10 epochs with a batch size of 32, learning rate of 2e-5, and AdamW optimizer. The XGBoost difficulty regressor used 500 estimators with a maximum depth of 6 and a learning rate of 0.05. All experiments were conducted on a server equipped with NVIDIA A100 GPU (80 GB), Intel Xeon Gold 6238R CPU, and 256 GB RAM. Training and evaluation code is implemented in Python using HuggingFace Transformers, Scikit-learn, and XGBoost libraries.

VI.RESULTS AND DISCUSSION

6.1 Cognitive Level Classification

The fine-tuned BERT classifier achieved a macro-averaged F1-score of 0.89 and an accuracy of 91.4% on the held-out test set, substantially outperforming the keyword-matching baseline (F1: 0.61) and the TF-IDF+SVM baseline (F1: 0.74). Error analysis revealed that the most common misclassifications occurred between adjacent cognitive levels (e.g., "Understand" vs. "Apply"), which is consistent with findings reported in prior literature and reflects genuine linguistic ambiguity at the boundaries of taxonomic categories. The highest accuracy was observed for the "Remember" level (95.2%), as questions at this level tend to feature highly distinctive linguistic markers such as "define", "list", and "state".

6.2 Difficulty Estimation

The XGBoost difficulty regressor achieved a MAE of 0.08 on the 0-1 normalized difficulty scale and a Pearson correlation of 0.91 with expert-assigned difficulty ratings. The deep learning baseline (fine-tuned RoBERTa) achieved a slightly lower correlation of 0.87, suggesting that the explicitly engineered features (lexical complexity, syntactic depth, semantic density) provide meaningful signals beyond those captured by pre-trained language representations alone. Feature importance analysis identified semantic density and parse tree depth as the two most predictive features.

6.3 Linguistic Quality Analysis

The linguistic quality module detected grammatical and stylistic issues with a precision of 0.86 and recall of 0.81 against expert-identified issues. Ambiguity detection achieved precision of 0.79, a more challenging task. Readability scores were highly consistent with expert judgments (Spearman rho = 0.88), validating the use of established readability formulae for this domain. The module identified a total of 2,341 issues across the test set, an average of 11.7 per paper, providing actionable feedback for paper revision.

6.4 Syllabus Coverage Evaluation

The syllabus coverage module correctly identified topic presence with an average precision of 0.84 and recall of 0.82 across all six disciplines. Coverage evaluation revealed that, on average, question papers in the dataset covered 73.4% of the identified syllabus topics, with significant variation across disciplines (range: 61.2% - 88.7%). Computer Science papers showed the highest coverage, while History papers demonstrated the most uneven topic distribution, a finding consistent with the broader and more interpretive nature of historical curricula.

6.5 Overall Quality Scoring

The composite AQPQA score correlated strongly with holistic expert quality ratings (Pearson r = 0.93, Spearman rho = 0.91), indicating that the weighted aggregation of module scores effectively captures the human notion of question paper quality. The framework significantly outperformed the GPT-3.5 zero-shot baseline (r = 0.74), demonstrating the value of specialized, multi-module architectures over general-purpose language models for this task. Processing time averaged 18.3 seconds per question paper, enabling practical real-time use by examination boards.

6.6 Comparison with Baselines

Across all evaluated metrics, the proposed AQPQA framework consistently outperformed all four baseline methods. The improvements were most pronounced for cognitive level classification and overall quality scoring, where the structured multi-module approach provided clear advantages over single-model and zero-shot alternatives. These results validate the design decision to employ modular, task-specific models rather than a monolithic end-to-end architecture.

VII.IMPLICATIONS AND LIMITATIONS

7.1 Practical Implications

The AQPQA framework offers substantial practical value for educational institutions, examination boards, and faculty members. By automating time-consuming quality review processes, the system can reduce the workload on expert reviewers and accelerate the question paper approval cycle. The detailed, dimension-specific feedback generated by the system provides actionable guidance for paper revision, supporting iterative improvement. The framework is particularly valuable for institutions

with limited access to examination review expertise, enabling consistent quality standards across decentralized or resource-constrained settings.

At the policy level, adoption of AI-based assessment tools can contribute to greater fairness and equity in educational assessment by reducing the influence of individual reviewer biases and institutional inconsistencies. The system's modular design allows institutions to customize the evaluation criteria and weight assignments to align with their specific quality frameworks and accreditation requirements.

7.2 Limitations

Several limitations of the current study warrant acknowledgment. First, the training corpus, while diverse, is limited to ten institutions and six disciplines, potentially restricting the generalizability of the models to other educational contexts, particularly non-English-language assessments or highly specialized technical fields. Second, the cognitive level classifier operates at the individual question level and does not account for interactions between questions or the cumulative cognitive demands of the paper as a whole. Third, difficulty estimation relies partly on historical student performance data, which may encode existing biases and structural inequities in the educational system.

Fourth, while the system provides interpretable module-level scores, the internal workings of the deep learning components remain partially opaque, which may limit acceptance by faculty and administrators who prefer fully transparent evaluation criteria. Fifth, the current implementation does not handle assessment formats beyond written examination papers, excluding oral examinations, practical assessments, and project-based evaluations from its scope.

VIII.CONCLUSION AND FUTURE WORK

This paper presented AQPQA, a comprehensive AI-based framework for automated quality assessment of examination question papers. The framework integrates transformer-based NLP models, machine learning regressors, linguistic analysis tools, and ontology-driven content mapping to evaluate question papers across five key quality dimensions. Experimental evaluation on a corpus of 1,200 question papers demonstrated strong performance across all modules, with the composite quality score achieving a Pearson correlation of 0.93 with human expert ratings.

The findings demonstrate that AI-driven tools can provide scalable, consistent, and actionable quality feedback for question papers, addressing a significant operational challenge for educational institutions. The modular architecture of AQPQA ensures adaptability to diverse assessment contexts and institutional requirements, while the interpretable output format supports transparent and evidence-based decision-making by educators.

Future work will pursue several directions. First, the framework will be extended to support multilingual question papers, incorporating multilingual transformer models such as mBERT and XLM-R. Second, the cognitive level classifier will be enhanced to capture inter-question relationships and paper-level cognitive profiles. Third, adversarial robustness testing will be conducted to evaluate system performance against edge cases and unusual question formats. Fourth, a longitudinal study will examine the impact of AQPQA adoption on the quality of examination papers and student outcomes in participating institutions. Finally, efforts will be directed toward improving model explainability through attention visualization and SHAP-based feature attribution, supporting greater stakeholder trust and adoption.

Acknowledgements

The authors gratefully acknowledge the participating institutions for providing examination data, the educational specialists who contributed expert annotations, and the anonymous reviewers for their constructive feedback. This research was supported in part by [Funding Body] under Grant No. [XXXX].

REFERENCES

1. Anderson, L. W., & Krathwohl, D. R. (2001). A taxonomy for learning, teaching, and assessing: A revision of Bloom's educational objectives. Longman.
2. Alsubait, T., Parsia, B., & Sattler, U. (2016). Measuring similarity in ontologies: A new family of measures. International Journal of Artificial Intelligence in Education, 26(1), 59-100.
3. Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater v.2. Journal of Technology, Learning, and Assessment, 4(3).
4. Benedetto, L., Crammer, K., & Donat, A. (2021). R2DE: A NLP approach to estimating IRT parameters of newly generated questions. Proceedings of the 11th International Conference on Learning Analytics and Knowledge (LAK'21), 361-370.
5. Denny, P., Luxton-Reilly, A., & Tempero, E. (2008). All syntax errors are not equal: Towards a classification of novice programming errors. Proceedings of the 13th Annual Conference on Innovation and Technology in Computer Science Education (ITiCSE), 346-350.
6. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT 2019, 4171-4186.
7. Foltz, P. W., Laham, D., & Landauer, T. K. (1999). Automated essay scoring: Applications to educational technology. Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications, 939-944.
8. Gierl, M. J., Lai, H., Hogan, J., & Matovinovic, D. (2017). A method for generating educational test items that are aligned to the common core state standards. Journal of Applied Testing Technology, 15(1), 1-18.
9. Heilman, M., & Smith, N. A. (2010). Good question! Statistical ranking for question generation. Proceedings of NAACL-HLT 2010, 609-617.
10. Huang, Z., Liu, Q., Chen, E., Zhao, H., Gao, M., Wei, S., Su, Y., & Hu, G. (2017). Question difficulty prediction for reading problems in standard tests. Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI-17), 1352-1359.
11. Lalor, J. P., Wu, H., & Yu, H. (2019). Learning latent parameters without human response patterns: Item response theory with artificial crowds. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), 4240-4250.
12. Li, X., Zhang, H., & Zhou, Y. (2022). Topic-modelling-based syllabus coverage analysis for automated examination evaluation. Computers & Education, 178, 104405.

13. Mitkov, R., & Ha, L. A. (2003). Computer-aided generation of multiple-choice tests. Proceedings of the HLT-NAACL 2003 Workshop on Building Educational Applications Using NLP, 17-22.
14. Noy, N. F., & McGuinness, D. L. (2001). Ontology development 101: A guide to creating your first ontology. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05.
15. Page, E. B. (1966). The imminence of grading essays by computer. Phi Delta Kappan, 47(5), 238-243.
16. Peng, J., Wang, L., & Zhao, Y. (2023). Holistic examination quality assessment using multi-dimensional AI evaluation: A large-scale empirical study. Expert Systems with Applications, 214, 119151.
17. Rodriguez, C., Gutierrez, F., & Deco, C. (2021). Knowledge graph-based curriculum alignment and examination coverage evaluation. IEEE Transactions on Learning Technologies, 14(4), 501-514.
18. Settles, B., LaFlair, G. T., & Hagiwara, M. (2020). Machine learning–driven language assessment. Transactions of the Association for Computational Linguistics, 8, 247-263.
19. Tsangaratos, P., Ilia, I., & Loupasakis, C. (2020). Automated mapping of examination learning outcomes to Bloom's Taxonomy using transfer learning. Applied Sciences, 10(21), 7571.
20. Yahya, A. A., & Osman, A. (2012). Automatic classification of questions in Bloom's taxonomy based on question structure. International Journal of Engineering Research and Technology, 1(3), 1-6.