International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2025: 7.089

QEvalAI- LLM Based Question Paper Quality Checker and Answer Evaluation System

Sneha S Krishnan[1], Bindu B[2]

[1]Department of Computer Applications, Musaliar College of Engineering & Technology, Pathanamthitta, Kerala, India
Email: *snehasksrd[at]gmail.com*

[2]Professor, Department of Computer Applications, Musaliar College of Engineering & Technology, Pathanamthitta, Kerala, India

Abstract: ***The academic process has a key- parts and two of the most important ones are examination evaluation and question paper preparation. The old ways of evaluating exams are slow not fair and can be influenced by how tired the person grading is or what they think about certain things. This paper is about QEvalAI, a system that uses Artificial Intelligence to check the quality of question papers and grade answers. The system uses a different tool, like Natural Language Processing, Machine Learning, Large Language Models and Optical Character Recognition to understand what the answers really mean rather than just looking for certain words. By making the evaluation process more accurate, fair and efficient QEvalAI helps teachers have work to do and it makes the whole academic assessment process better. QEvalAI is a step, towards making evaluation systems smarter and more automated which is what modern education needs.***

Keywords: QEvalAI, LLM, NLP, OCR, Answer Evaluation, Bloom's Taxonomy

1. Introduction

You know exams are really important in checking how well students are doing in school and what they've learned. The way we prepare question papers and check answer sheets can make a difference in how fair and good the education is. Usually making question papers and checking answers is done by hand, which takes a lot of time and can be unfair because people might make mistakes or be biased.

Most computer systems that try to help with this use techniques like looking for specific words. These systems can't really understand if a student gets the idea or not or if they can think logically and make sense. They also don't check if the questions are good and fair or if they're too easy or too hard.

Now with new tech, like Artificial Intelligence, big language models and computer vision we can make systems that really understand what people are saying. These systems can get the meaning behind the words see if someone is just copying and check if they're thinking logically like a teacher would.

These new systems also give feedback and help teachers see how students are doing so they can teach better.

The system can even read answers that are written or typed and it gives teachers information away so they can help students learn better.

This helps teachers see where students are struggling and make their teaching better.

In addition to improving evaluation accuracy, the proposed system also focuses on enhancing the overall quality of question papers. It ensures that questions are well-structured, balanced in difficulty level, and aligned with learning objectives such as Bloom's Taxonomy. This helps in creating fair and effective assessments for students. Ultimately, the system is designed to assist human teachers by reducing their workload while ensuring fair and accurate evaluation.

2. Related Works

Recent advancements in intelligence have greatly impacted academic assessment and examination systems. Researchers have been working on developing systems that can automate question paper generation improve answer evaluation accuracy and enhance overall academic quality. The use of techniques like Large Language Models, Natural Language Processing and Optical Character Recognition has enabled systems to move beyond traditional keyword-based approaches and focus on semantic understanding and contextual evaluation.

One study, Automatic Question Generation using Large Language Models (2024) looked into using intelligence to generate structured questions from educational materials. The study showed that Large Language Models can produce types of questions with varying difficulty levels. This can help reduce teacher workload and support assessment creation. However ensuring accuracy may still require manual validation.

Another study, Bloom's Taxonomy-based Question Classification using NLP Techniques (2023) focused on classifying examination questions into cognitive levels. This approach helps maintain balance in question papers and improves academic standards. However, the study noted challenges in handling question phrasing and complex subject-specific terminology.

There is also Automated Grammar and Clarity Checking in Educational Assessments (2023) which introduced systems that use NLP models to identify errors and improve question clarity. These systems enhance readability and fairness in examinations. However, they may sometimes produce suggestions when dealing with domain-specific terms.

The study AI-Assisted Examination Systems for Higher

Education (2024) investigated the use of AI in automating question generation, grading and performance analysis. The study highlighted improvements in efficiency and consistency. It emphasized that AI should complement than completely replace educators.

Lastly Evaluation of AI-Based Question Paper Quality Analysis Systems (2024) examined systems that assess question paper quality based on parameters like clarity, difficulty distribution and cognitive balance. The research concluded that combining AI analysis with expert review provides the reliable and accurate results.

3. Outlined Method

The QEvalAI system is being designed to automate evaluation processes and improve the quality of question papers and answer assessment. The system combines intelligence techniques, web technologies and database systems to provide an efficient and intelligent evaluation platform.

3.1 Requirement Analysis

The requirement analysis phase focuses on identifying the limitations of examination systems. Manual question paper preparation and answer evaluation are time-consuming and prone to bias. They often lead to inconsistencies in grading. Existing automated systems mainly rely on keyword-based evaluation. This fails to assess understanding, reasoning ability and contextual correctness of descriptive answers. There is also no mechanism to evaluate question paper quality in terms of Bloom's Taxonomy, difficulty distribution and redundancy.

To address these challenges the system defines functional requirements. These include automated question paper generation, Bloom's Taxonomy classification, difficulty level analysis OCR-based handwritten answer processing, semantic answer evaluation, plagiarism detection and performance analytics. Non-functional requirements include system usability, scalability, accuracy, performance efficiency and secure data management.

a) System Design

The QEvalAI system architecture is designed as a structure. Different components interact with each other through a database. The major modules of the system include:
labelsep

- **Admin Module:** This manages the operation of the system. It includes user management, monitoring system activities and maintaining system data.
- **Teacher Module:** This allows teachers to generate question papers. They can also analyze question paper quality upload student answers and view evaluation results and reports.
- **Question Paper Generation Module:** This uses Large Language Models to generate question papers from uploaded syllabus materials. It ensures format and marks distribution.
- **Question Paper Quality Analysis Module:** This evaluates question papers based on Bloom's Taxonomy levels, difficulty classification, grammar checking and redundancy detection.
- **Answer Evaluation Module:** This performs evaluation of student answers using Natural Language Processing techniques. It assigns marks based on understanding.
- **OCR Processing Module:** This converts. Scanned answer sheets into machine-readable text for further evaluation.
- **Plagiarism Detection Module:** This identifies copied or similar answers using text similarity and semantic comparison techniques.
- **Reports and Analytics Module:** This generates reports. These include marks, feedback, Bloom's taxonomy distribution and student performance analysis.

All these modules are interconnected. They communicate with a database system that stores user data, question papers, student answers, evaluation results and analytical reports. This ensures data management and smooth system operation.

b) Development

The system is developed using technologies to ensure efficiency and scalability. The backend is implemented using Python and the Django framework. This handles business logic, API services and database interactions. The frontend is developed using HTML, CSS and JavaScript. This provides a user- interface. MySQL is used as the database management system. It. Manages all application data efficiently. Artificial intelligence techniques, like Natural Language Processing are used for semantic answer evaluation and question analysis. Optical Character Recognition is applied to process answer sheets. Large Language Models are used to understand context evaluate answers and generate feedback.

c) Integration & Testing

After development all modules are integrated into a system. This ensures communication and functionality. Integration testing is performed to verify that all components work together without errors. Functional testing is conducted to validate features. These include question paper generation, answer evaluation, plagiarism detection and report generation. Performance testing ensures that the system operates efficiently under conditions. Usability testing evaluates the ease of use and user interaction. These testing processes help. Resolve issues. They ensure that the final system is reliable, accurate and user-friendly.

4. Evaluation & Optimization

The QEvalAI system needs to be checked to see how well it is working. This means looking at all the parts of the system. We need to check how well the artificial intelligence is generating questions. We also need to check how well the system is understanding the answers. The QEvalAI system has to be able to read answers correctly. It also has to be able to find out if someone is copying someone Work. We check the system to see how accurate it is how long it takes to respond if the content is relevant and if the users are happy.

The part of the system that checks answers has to be correct and consistent. The part that reads answers has to be able to get the text right. The part that checks for copying has to be able to find work. To make the system better we use

techniques like improving the natural language processing making the large language models better and making the database work faster. This makes the system more efficient and reliable.

5. Proposed Methodology

The QEvalAI system has a plan to automatically check answers and make assessments better. The system takes the answers given by students. Does several things to them. First it gets the answers ready. Then it takes out the features. After that it compares the answers to see if they are similar. The system then gives a score. Provides feedback. The QEvalAI system is designed to make the process of checking answers more accurate.

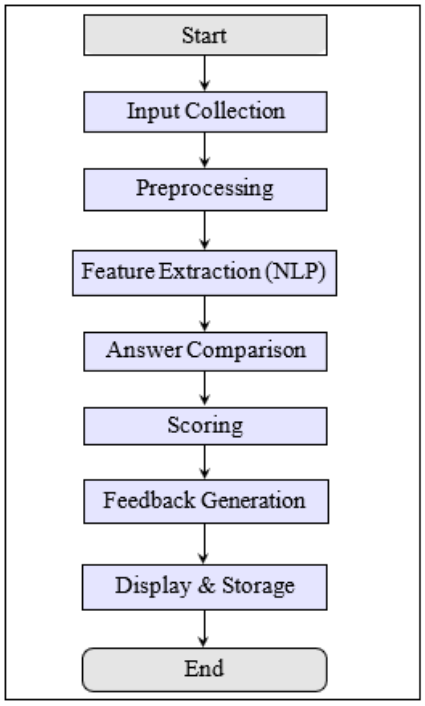

Figure 1: Workflow of QEvalAI Answer Evaluation

The process starts with collecting information. The system gets the question paper and the answer key. What the students wrote. Then the system makes sure the information is clean and easy to use. If the students wrote their answers by hand the system uses a tool to read the writing. The system looks for things, like keywords and what the words really mean. The student answer is compared to the answer to see how similar they are. The system gives marks based on how correct and complete the student answer's The system tells the student what they did well and what they need to work on. The question paper and the student answers and the results are all stored in the database for the system to use later. The system shows the results to the students and the teachers.

6. System Architecture

The QEvalAI system is built with a design. This helps it to be modular, scalable and good at handling data. Each layer does a job to turn raw input into useful evaluation results.

The input layer gathers data from users. This includes question papers and student answers. The preprocessing layer cleans the data. Uses OCR to turn handwritten text into digital format. The feature extraction layer uses Natural Language Processing. It extracts features from the text. The AI evaluation layer uses Large Language Models. It. Compares answers giving scores and feedback. The output layer shows evaluation results. It also saves them in the database.

7. Result & Discussion

1) System Performance

The QEvalAI system works well. It automates evaluation and improves assessment quality. It evaluates answers. It uses understanding, not just keywords. The system uses AI techniques like Natural Language Processing and Large Language Models. It gives scores and feedback with little human help.

The OCR module converts handwritten answers into text. It handles both typed and scanned inputs. The extracted text is refined using AI. This improves clarity and correctness. The answer evaluation module compares student answers with reference answers. It ensures consistent grading.

The QEvalAI systems performance is good. It uses Python, Django and MySQL. These technologies help with data handling, fast processing and smooth user interaction. The database structure is centralized. It makes it easy to store and retrieve question papers, student answers and evaluation results. This makes the system reliable and scalable. The QEvalAI system and its Large Language Models are key, to its success. The QEvalAI systems performance is enhanced by its design and technologies.

2) Test Cases and Outcomes

The system was tested in different ways to see how well it works and if it is reliable. We did a lot of tests to check the parts of the system.

The part that generates questions worked well. It made question papers based on the material we uploaded. The part that checks answers also worked well. It looked at what the students wrote. Gave them the right scores and feedback. The part that converts answers into text worked too but sometimes it had a little trouble if the handwriting was not clear. We can fix this when we get the answers ready to be looked at.

We also tested the part that checks for plagiarism. It looked at answers from students and found the ones that were similar or copied. The part that makes reports and analyzes data gave us reports on how the students did including their scores, feedback and how well they performed.

We tested the system on devices and in different environments to make sure it works properly and is easy to use. The test results show that the system works well and gives accurate results. The results also indicate that the system maintains consistency in evaluation across different test cases, reducing variations caused by manual checking. Overall, the testing confirms that the system is reliable,

efficient, and suitable for real-world academic applications.

8. Comparative Analysis with Existing Systems

When we compare QEvalAI to exam systems we can see that it is much better in many ways. The old way of grading exams is slow and not always fair. It relies on people checking the answers by hand and looking for keywords, which does not really check if the student understands the material. This way is also prone to bias and mistakes.

QEvalAI is different. It uses intelligence to understand the answers and give grades that are accurate and consistent. It can look at the context. Evaluate the students reasoning, which is not possible with the old way.

QEvalAI is also a system that does everything from generating question papers to analyzing performance. This makes it easier for teachers to do their jobs and reduces their workload.

The system can even handle answer sheets, which makes it easy to switch from traditional exams to digital exams. All the data is stored in one place, which makes it easy to organize and access.

Overall QEvalAI shows that using intelligence, in academic evaluation is a good idea. It makes the process of grading exams more accurate and more intelligent.

Model Evaluation using ROC Curve

To see how well the answer evaluation model works we use something called a ROC Curve. The ROC Curve shows us how the model does in terms of incorrect answers. It helps us understand how good the model is at telling answers from incorrect ones.

The True Positive Rate is the number of answers that the model gets right. The False Positive Rate is the number of answers that the model thinks are correct. A good model will have a curve that's close to the top left corner. This means the model is very good, at evaluating answers. The ROC Curve is a way to evaluate the answer evaluation model. The answer evaluation model is what we are trying to make.

The Area Under the Curve or AUC is something we use to see how well something is working. If the AUC value is close to

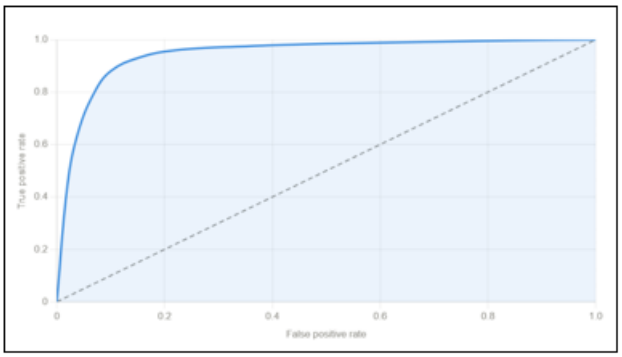

Figure 2: ROC Curve for QEvalAI Answer Evaluation Model

1 that means the system is very accurate and good at what it does.Thus, the ROC curve and AUC value together provide a clear measure of the model's performance and reliability in answer evaluation.

9. Conclusion

QEvalAI is a good solution for evaluating academic work. It uses intelligence to look at question papers and answers. This system uses things like Natural Language Processing and Large Language Models to understand what the answers mean than just looking for certain words. This way we do not need people to do all the work. We can grade things more accurately.

QEvalAI can handle types of answers like typed and handwritten ones. This makes it easier for students and teachers to use. The system can even turn answers into text that computers can read. Then it uses Natural Language Processing and Large Language Models to evaluate the answers and give feedback.

The QEvalAI system works well for things like making question papers checking quality evaluating answers finding plagiarism and looking at performance. It uses technologies like Python and MySQL to handle data and make sure the system runs smoothly.

Overall QEvalAI shows us that using intelligence in academic evaluation can make things more efficient, fair and consistent. This system is a step towards making examination systems smarter and more automated. It can help make education platforms that use intelligence even better in the future.

In the future we can make QEvalAI even better by adding things like learning and the ability to evaluate answers in many languages. We can also make the system work better with education platforms. If we can make the models more accurate and give feedback in time that will make the system even more useful for users and help them learn more. As artificial intelligence gets better systems, like QEvalAI can help make education more personalized and smarter. Even though the system improves evaluation, human involvement is still important to ensure fairness and final judgment. With further improvements, QEvalAI can become more accurate, reliable, and useful in real-world educational environments. Overall, this work shows how technology can support education in a practical and meaningful way.

References

[1] T. Brown et al., "Language Models are Few-Shot Learners," *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
[2] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of NAACL-HLT*, 2019.
[3] A. Radford et al., "Language Models are Unsupervised Multitask Learners," *OpenAI Research Report*, 2019.
[4] D. Jurafsky and J. H. Martin, "Speech and Language Processing," *Pearson Education*, 2021.
[5] A. Vaswani et al., "Attention is All You Need," *Advances in Neural Information Processing Systems*

(NeurIPS), 2017.
[6] R. Smith, "An Overview of the Tesseract OCR Engine,"
[7] *Proceedings of ICDAR*, 2007.
[8] S. Kumar and P. Sharma, "AI-Based Automatic Question Generation for Educational Assessment," *International Journal of Educational Technology*, 2022.
[9] A. Singh and D. Verma, "Automated Evaluation of Descriptive Answers using NLP," *International Journal of Artificial Intelligence in Education*, 2023.
[10] L. Zhao and P. Wang, "Semantic Answer Evaluation using Deep Learning Models," *Journal of Intelligent Systems*, 2023.
[11] Y. Chen and H. Li, "AI-based Academic Integrity and Plagiarism Detection Systems," *IEEE Access*, 2024.