International Journal of Enhanced Research in Educational Development, IJERED,
ISSN: 2320-8708, Vol. 13 Issue 2, Mar.-Apr, 2025

# AQVCA: A Hybrid AI-based Approach for Automatic Question Validation and Classification in Educational Assessments

Monika Singh[1], C. Patvardhan[2], C. Vasantha Laxshmi[3]

[1,2,3]Dayalbagh Educational Institute, Agra, UP, India

**ABSTRACT**

**Ensuring high-quality questions in educational assessments is essential for effective learning and evaluation. This paper introduces AQVCA (Automated Question Validation and Classification Algorithm), an advanced AI-driven framework that refines question processing through syntactic validation, semantic coherence analysis, duplicate detection using SBERT embeddings, and Bloom's Taxonomy classification. Our approach eliminates grammatically incorrect, semantically misleading, and redundant questions, ensuring only well-structured and meaningful queries are retained. The system surpasses traditional NLP-based and deep learning-based models, achieving superior precision, recall, and F1-score on EduQA and TREC-QA datasets. Additionally, its seamless integration into intelligent tutoring systems and online assessments enhances automated evaluation with minimal manual intervention. The proposed method significantly improves question quality, making it a scalable, efficient, and intelligent solution for modern educational assessments.**

**Keywords: Automatic Question Validation, Natural Language Processing, Bloom's Taxonomy Classification, Semantic Coherence Analysis, Educational Assessment**

## INTRODUCTION

The rapid evolution of digital education and e-learning platforms has transformed the way students acquire knowledge and how educators design assessments. With the proliferation of online courses, Learning Management Systems (LMS), and AI-powered tutoring systems, the need for automated question validation and classification has become more critical than ever. Well-structured, semantically correct, and pedagogically relevant questions are the backbone of effective assessments (Paiva et al., 2022). However, ensuring the quality of assessment questions remains a major challenge due to issues such as syntactic errors, conceptual inaccuracies, duplicate questions, and poor classification of cognitive complexity.

The manual curation and validation of assessment questions by educators is a time-consuming, inconsistent, and error-prone process that struggles to scale with the increasing volume of digital content. While rule-based automated systems attempt to address this challenge, they often lack the ability to handle semantic variations and deep contextual understanding (Gašpar et al., 2023). Machine learning (ML) and deep learning (DL) models provide promising solutions but still face limitations in domain adaptability, duplicate detection, and structured classification (Zhai et al., 2020) (Süzen et al., 2020). Key challenges include grammatical errors, logically inconsistent or misleading questions, and redundancy due to ineffective keyword-based similarity detection. Additionally, existing classification methods lack the granularity needed for accurate Bloom's Taxonomy-based categorization. Many solutions rely on human intervention, making large-scale automation difficult. Furthermore, general AI models, such as ChatGPT, often generate answers for poorly formed questions rather than filtering them, leading to misinformation (Maxnun et al., 2024). Addressing these issues requires a robust AI-driven system for efficient, scalable, and high-quality question validation and classification.

Given the limitations of existing approaches, there is a critical need for an advanced AI-powered question validation and classification system that ensures high-quality, error-free, and pedagogically sound questions for educational assessments. Such a system should automate question validation and filtering to reduce human effort and increase efficiency. Additionally, it should incorporate semantic understanding and context-aware processing for superior question evaluation, offer accurate Bloom's Taxonomy-based classification to support structured assessments, and ensure scalability to handle large question banks across different domains and languages (Revanesh et al., 2023; Sailer et al., 2021).

To overcome the limitations of traditional question validation systems, we introduce AQVCA (Automated Question Validation and Classification Algorithm), an advanced NLP-driven framework that enhances question validation, filtering, and classification. AQVCA integrates Natural Language Processing (NLP), machine learning (ML), and deep learning (DL) techniques to ensure robust question assessment. Unlike rule-based or keyword-based methods, AQVCA employs NLP-based grammatical analysis to eliminate structurally incorrect questions while leveraging SBERT-based embeddings, Named Entity Recognition (NER), and dependency parsing for semantic coherence. Redundant questions are filtered using vector embedding techniques, and Bloom's Taxonomy-based classification is achieved through Transformer-based models. AQVCA surpasses existing techniques by combining rule-based interpretability, ML precision, and DL adaptability, addressing challenges in semantic variation, duplicate detection, and structured classification. By merging linguistic and deep learning approaches, AQVCA ensures accurate, scalable, and intelligent question processing, making it a significant advancement in automated assessment management.

The implementation of AQVCA has far-reaching implications in the education sector and beyond. It enhances the quality of assessment questions by ensuring that students encounter only well-structured, meaningful, and pedagogically appropriate questions. The system also improves scalability in online learning, making it easier to integrate into e-learning platforms, MOOCs, and intelligent tutoring systems (Ariely et al., 2023). By automating the validation process, AQVCA reduces the burden on educators and assessment designers, enabling them to focus on content creation and pedagogy. Additionally, the system supports AI-powered adaptive learning by generating personalized question sets tailored to students' learning progress. AQVCA can also be integrated into computer-based testing (CBT) platforms and AI-driven exam systems, improving test security and question reliability. Table 1 shows the academic purposes of the AQVCA system. Beyond education, this system can be applied to AI-powered hiring assessments, corporate training modules, and knowledge evaluation systems in various industries(Bird et al., 2023).

**Table 1: Academic purposes of an AQVCA System**

| Application | Purpose |
|---|---|
| **Question Validation** | Ensures grammatical and logical correctness. |
| **Duplicate Detection** | Removes redundant or similar questions (Koswatte & Hettiarachchi, 2021). |
| **Bloom's Classification** | Categorizes questions by cognitive levels (Patil et al., 2022). |
| **AI Question Generation** | Creates high-quality exam questions. |
| **Adaptive Learning** | Recommends questions based on student progress. |
| **Plagiarism & Bias Check** | Detects reused or biased content. |
| **Real-Time Validation** | Validates questions instantly in online exams. |
| **Smart Question Bank** | Organizes and manages large question sets. |

Thus, the objectives of the paper are summarized as:

- Provide an approach to automatically remove syntactically and semantically incorrect and redundant questions from a large question-bank.
- Optimize and classify questions by categorizing them using Bloom's Taxonomy.
- Enhance AI-driven learning and evaluation with a standardized, error-free question bank.

The rest of the paper is organized as follows. Section 2 presents the six levels of Bloom's Taxonomy categories with respective keywords. Section 3 discusses various question-generation approaches. Section 4 presents the proposed approach and techniques to achieve the objectives. Section 5 presents the experimental results of our approach on a large dataset along with an example. Section 6 provides the conclusion.

## BLOOM'S TAXONOMY

In 1956, Benjamin S. Bloom, a professor at theUniversity of Chicago, presented a document that contained a taxonomy known as Bloom's Taxonomy to categories the educational goals, outlining the three domains of learning as follows: cognitive, psychomotor, and affective domains. Knowledge, Comprehension, Application, Analysis, Synthesis, and Evaluation are the six key categories in the cognitive domain that are covered by the original Bloom's Taxonomy. The categories were arranged from straightforward to intricate and from tangible to abstract: lower-order abilities that require less cognitive processing to higher-order talents that require more in-depth study and cognitive processing (Bhanuprakash & Jayaram, 2022).

Revision of Bloom's Taxonomy: A new document by David R. Krathwohl, was released in 2002. The updated taxonomy has two dimensions: cognitive processes and knowledge (Sucipto et al., 2024). The six initial Taxonomy categories were renamed as remember, understand, apply, analyze, evaluate, and create (Asthana et al., 2024) (Fig. 1).

These categories are briefly explained as follow:

1) *Remember*: Utilizing long-term memory to retrieve pertinent information, e.g., Recognizing (or identifying) and Recalling (or retrieving).
2) *Understand:* Describing the concept about what is being communicated in instructional communications, including those that are written, spoken, and visual description. E.g., Interpreting, Exemplifying, Classifying, Summarizing, Inferring, Comparing and Explaining.
3) *Apply:* Executing or applying a process in a circumstance. E.g., Carrying out (or completing) and Implementing (or using).
4) *Analyze:* Breaking a piece of content into its component parts and figuring out how those parts connect to each other and to a larger structure or goal. E.g., Differentiating, Organizing and Attributing.
5) *Evaluate:* Making judgements based on norms and criteria. E.g., Checking and Critiquing.
6) *Create:* Combining components to create a brand-new, cohesive whole or a unique product. E.g., Generating, Planning and Producing (Scaria et al., 2024).

**Table 2: Bloom's Taxonomy Verbs**

| **Level** | **Verbs** | | | |
|---|---|---|---|---|
| | Knowledge Dimensions | | | |
| | Factual | Conceptual | Procedural | Metacognitive |
| **Creating** | Generate<br>Write<br>Combine | Gather<br>Device<br>Plan | Design<br>Develop<br>Compose | Produce<br>Create<br>Actualize |
| | Invent, Categorize, Compile, Compose, Explain, Modify, Organize, Plan, Arrange, Summarize, Tell, Build, Choose, Construct, Estimate, Formulate, Imagine, Invent, Make-up, Originate, Predict, Propose, Solve, Discuss, Modify, Change, Improve, Adapt, Minimize, Maximize, Elaborate, Test, Improve | | | |
| **Evaluating** | Check<br>Criticize<br>Rank | Define<br>Review<br>Assess | Judge<br>Evaluate<br>Conclude | Reflect<br>Rate<br>Prioritize |
| | Appraise, Compare, Conclude, Defend, Describe, Discriminate, Explain, Justify, Relate, Summarize, Support, Award, Decide, Determine, Dispute, Measure, Mark, Recommend, Select, Agree, Prove, Perceive, Value, Estimate, Influence, Deduct | | | |
| **Analyzing** | Choose<br>Classify<br>Order | Distinguish<br>Identify<br>Explain | Integrate<br>Compare<br>Differentiate | Match<br>Analyze<br>Achieve |
| | Break Down, Contrast, Deconstruct, Illustrate, Infer, Outline, Select, Separate, Categorize, Discover, Dissect, Divide, Examine, Inspect, Simplify, Survey, List, Assume, Conclude | | | |
| **Applying** | Use<br>Answer<br>Classify | Give<br>Set<br>Experiment | Carry out<br>Employ<br>Calculate | Select<br>Enhance<br>Execute |
| | Apply, Change, Compute, Construct, Demonstrate, Manipulate, Modify, Operate, Predict, Prepare, Produce, Show, Solved, Build, Choose, Develop, Interview, Make-Use, Organize, Experiment, Plan, Utilize, Model, Identify | | | |
| **Understanding** | Interpret<br>Categorize<br>Summarize | Categorize<br>Describe<br>Consider | Paraphrase<br>Classify<br>Predict | Foresee<br>Explain<br>Execute |
| | Comprehend, Convert, Distinguish, Estimate, Extend, Generalize, Translate, Compare, Contrast, Demonstrate, Illustrate, Outline, Rephrase, Show, Classify, Infer, Exemplify, Tag, Comment, Annotate | | | |
| **Remembering** | Label<br>Spell<br>List | Recognize<br>Name<br>Consider | Recall<br>Recap<br>Tabulate | Outline<br>Identify<br>Omit |
| | Retrieve, State, Define, Know, Match, Reproduce, Select, Omit, Choose, Find, Show, Relate, Tell, Locate, Point-out, Highlight, Bookmark, Search. | | | |

Table 2 represents a certain level of cognitive domain in the knowledge dimension which gives an idea about how to ask questions to evaluate the knowledge of the students to assess their cognitive level (Higher or lower) (Chindukuri & Sivanesan, 2024; Jeslin Shanthamalar et al., 2024).

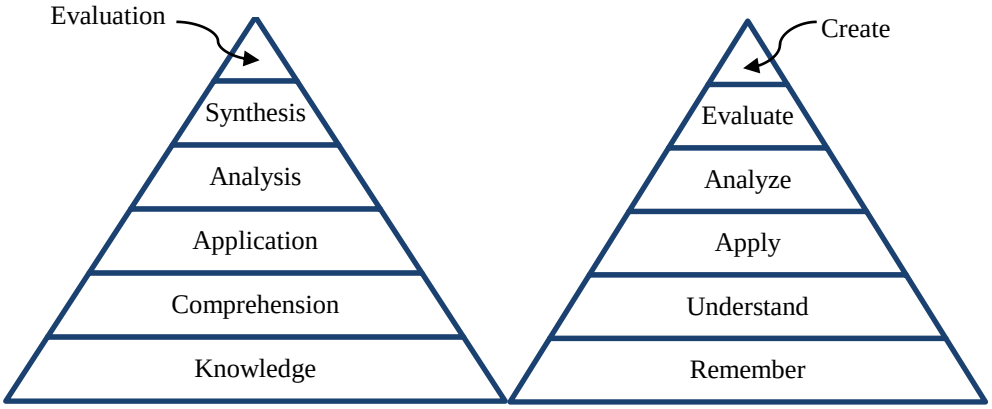


**Fig 1: Original and Revised Bloom's Taxonomy Levels**

## RELATED WORK

**Rule-Based and NLP-Based Filtering**Early automated question validation systems relied on rule-based methods, where predefined linguistic rules were used to filter grammatically incorrect or illogical questions. While effective for basic filtering, these systems were highly rigid and lacked adaptability to different linguistic variations (Gašpar et al., 2023). Later, NLP-based techniques such as syntactic parsing, part-of-speech (POS) tagging, and Named Entity Recognition (NER) improved filtering accuracy by analyzing sentence structure and coherence (Supriyono et al., 2024). However, rule-based and NLP methods still struggled with semantic variations and domain-specific adaptability, limiting their effectiveness for large-scale question validation.

**Machine Learning-Based Question Classification**
Supervised machine learning (ML) approaches, such as Support Vector Machines (SVM), Decision Trees, and Random Forests, have been employed for question validation and classification tasks. These models rely on handcrafted features, such as word embeddings, syntactic structures, and lexical similarities, to classify and filter questions (Tri et al., 2006). However, the reliance on manual feature engineering makes these models less scalable for complex, domain-specific educational datasets. Moreover, keyword-based matching techniques used in ML approaches often fail to detect semantic similarities between reworded questions, leading to redundancy in question banks (Hussan, 2020).

**Deep Learning-Based Approaches**
Deep learning (DL) techniques, particularly Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTMs), and Convolutional Neural Networks (CNNs), have demonstrated superior contextual understanding compared to traditional ML approaches (Sharma & Gupta, 2018). These models capture long-range dependencies in text, improving question classification and duplicate detection. However, they require large-scale labeled datasets and often struggle with generalization across different domains (Talaei Khoei et al., 2023).

Recent advancements in Transformer-based architectures, such as BERT, SBERT, and GPT-4, have significantly improved semantic understanding, duplicate detection, and contextual coherence in question validation (Barlybayev & Matkarimov, 2024). AQVCA leverages SBERT embeddings for semantic similarity analysis and Graph Neural Networks (GNNs) for advanced duplicate detection, overcoming the limitations of earlier models (Deena et al., 2024).

**Bloom's Taxonomy-Based Classification**
Bloom's Taxonomy is widely used in educational assessment to classify questions into six cognitive levels: Remember, Understand, Apply, Analyze, Evaluate, and Create (Spivey, 2007). Early classification methods relied on keyword-based mapping, which often resulted in misclassification due to ambiguous or multi-contextual keywords (Shaikh et al., 2021).

More recent approaches use Transformer-based hierarchical classification models, which analyze action verbs and sentence structures to provide a more precise classification of questions (Sayeed et al., 2024). AQVCA incorporates Bloom's Taxonomy-based classification using Transformer models, ensuring accurate and contextually relevant question categorization (Bhattacharyya, 2025; Dutta et al., 2024).

## AUTOMATIC QUESTION GENERATION APPROACHES

Question Generation (QG) methods can be divided into three classes: rule-based question generation, neural-based question generation, and transformer-based question generation. Transformer-based technique is the most recently developed method for question generation (Divate & Salgaonkar, 2017).

**Rule-based Question Generation:** Rule-based question generation (QG) approaches are methods used to automatically generate questions from text based on predefined linguistic and syntactic rules. These approaches depend on the specific patterns and structure of the language in the text to generate meaningful questions (Riloff & Thelen, 2000). The Rule-based question-generation approaches are further classified as:

**1. Template-Based Question Generation**
This method utilizes predefined templates or patterns for question generation. The templates include question structures with placeholders that are filled in with specific information from the text. For example, given a sentence like "The capital of France is Paris," a template might be "What is the capital of [Country]?" which would generate the question "What is the capital of France?"

**2. Syntactic Transformation-Based Generation**
This approach involves transforming the syntactic structure of a sentence for question generation. It typically involves identifying key components of a sentence (such as the subject, verb, and object) and then modifying these components to form a question. For example, from the sentence "John is reading a book," this approach might generate the question "What is John reading?"

**3. Linguistic Rule-Based Generation**
This method utilizes detailed linguistic rules, such as those related to parts of speech, dependency parsing, or semantic roles, for question generation. It involves deep analysis of the sentence to understand its meaning and structure before generating questions. For example, given the sentence "Marie Curie discovered radium," the linguistic rules might identify the action (discovered) and the object (radium) to generate the question "What did Marie Curie discover?"

**4. Semantic Role Labeling-Based Generation**
This method utilizes semantic role labeling (SRL), which identifies the roles of different entities in a sentence (e.g., agent, object, instrument) for question generation. The system understands who did what to whom and uses this understanding to generate questions. For example, from the sentence "The chef cooked a delicious meal," SRL might identify "chef" as the agent and "meal" as the object, leading to the question "What did the chef cook?"

**5. Frame-Based Question Generation**
This approach includes frame semantics, where the meaning of a word is understood within a broader structure or "frame" that includes various elements like the participants, actions, and circumstances related to the word. Questions are generated based on these frames. For example, give a sentence "The teacher explained the theory," the frame for "explaining" might include elements like "teacher" (explainer), "theory" (what is explained), leading to questions like "Who explained the theory?" or "What did the teacher explain?"

**6. Transformation Rules with Information Extraction**
This approach includes combining syntactic transformations with information extraction techniques. The system first extracts key information from the text (like entities, relations, and events) and transformation rules are applied for question generation. For example, given a sentence "The Nobel Prize was awarded to Rabindranath Tagore in 1913," the system extracts the key entities and event, then generates questions like "Who was awarded the Nobel Prize in 1913?"
All the strengths and weaknesses of the Rule-based AQG systems are summarized in Table 3.

**Neural-based Question Generation:** The neural-based question generation approach utilizes sequence-to-sequence learning algorithm to learn long-term dependencies for making predictions during questiongeneration by making use of LSTM and Recurrent Neural Network (RNN). Neural-based models make an effort to go beyond simple text memorization (Kim et al., 2019). Neural-based question generation involves the following steps:

- By offering several embeddings with a global vector model (GloVe model), preprocessing transforms a raw sentence into a neural-based encoder-friendly message.
- The information-rich sentence from the previous stage is processed by an encoder-decoder in the question-construction step. By maximizing the conditional log- likelihood of the anticipated question sequence, a sequence-to-sequence learning algorithm generates the subsequent token from the previously chosen tokens andan input text.
- Post-processing is used to polish output generated previously (Guan et al., 2023) (Kumar et al., 2019).

**Table 3: Strengths and Weakness of Rule-based AQG Systems**

| Techniques | Strengths | Weaknesses |
|---|---|---|
| **Template-Based Question Generation** | Simple and effective for specific domains where the templates are well-defined. | Limited by the scope of the predefined templates and struggles with generating questions for more complex or less structured text. |

| | | |
|---|---|---|
| **Syntactic Transformation-Based Generation** | It can generate a wide variety of questions by altering sentence structures, making it more flexible than template-based methods. | May struggle with complex sentences or ambiguous syntax, leading to less accurate questions. |
| **Linguistic Rule-Based Generation** | More accurate and sophisticated, capable of generating meaningful questions even from complex sentences. | Requires extensive linguistic knowledge and resources to implement and may be computationally expensive. |
| **Semantic Role Labeling-Based Generation** | Generates questions that are closely tied to the meaning of the sentence, leading to more contextually relevant questions. | The complexity of accurately labeling roles, especially in nuanced sentences, can limit effectiveness. |
| **Frame-Based Question Generation** | Highly context-aware and capable of producing questions that align well with the underlying meaning of the sentence. | Requires comprehensive frame definitions and can be challenging to apply in domains with less well-defined frames. |
| **Transformation Rules with Information Extraction** | It can generate precise questions by focusing on the most relevant information in the text. | Depends on the quality of the information extraction process, which can vary depending on the complexity of the text. |

**Transformer-based Question Generation:** Transformer-based models use an encoder and a decoder with an attention mechanism to convert one sentence's sequence into another. The transformer has more power to encode many linkages and nuances for each word because of its multi-head attention mechanism (Lopez et al., 2021). Transformer-based model employs a multi-head attention mechanism in threedifferent ways:

- Self-attention in encoder to pay attention in inputsequence.
- Self-attention in decoder to pay attention in targetsequence.
- Encoder-decoder-attention in the decoder by which thetarget sequence pays attention to input sequence.

The input embedding and positional embedding, which capture the meaning and placement of each word, are significant components of the transformer-based paradigm (Matsumori et al., 2023).

## METHODOLOGIES

In this study, we propose a novel Automated Question Validation and Classification Algorithm (AQVCA) designed to systematically filter, refine, and classify questions within an educational question bank. The algorithm is developed to eliminate syntactically and semantically incorrect questions, detect duplicates, and classify valid questions based on Bloom's Taxonomy. Unlike traditional rule-based or manual methods, our approach integrates Natural Language Processing (NLP), Machine Learning (ML), and Deep Learning (DL) techniques to enhance accuracy, efficiency, and scalability.

The proposed methodology follows a six-stage pipeline, data preprocessing, syntax and grammar checking, semantic validation, duplicate removal, Bloom's classification, and final storage, ensuring a structured and automated workflow. Each step is carefully designed to enhance question quality while minimizing human intervention. Our proposed algorithm for the described method is properly presented in Table 4.
Here, our objective is to compute $Q_{final}$ which represents the set of validated, unique, and Bloom's taxonomy-classified questions, ensuring high-quality question processing and storage.

$$Q_{final} = T(C(D(S(G(P(Q))))))$$

The first step, data preprocessing, involves cleaning and normalizing the input text to ensure consistency. This includes removing special characters, converting text to lowercase, performing stemming or lemmatization, and correcting spelling errors.

$$Q = \{q_1, q_2, \ldots, q_n\}$$
$$Q_{clean} = P(Q) = \{P(q_1), P(q_2), \ldots, P(q_n)\}$$

Where, Q be the set of input questions and P(q) includes text normalization, tokenization, stemming/lemmatization, and spell-checking. These operations improve the quality of the text and make it easier for NLP models to process. For example, a question like *"Whatt is the capital of UUnited States?"* would be corrected to *"What is the capital of the United States?"*.

**Table 4: Proposed Algorithm for the AQVCA System**

| |
|---|
| 1. START<br>2. INPUT: Question Bank (Q)<br><br>3. Step 1: Data Preprocessing |

```
4. FUNCTION Preprocess(Q):
5.     CLEAN text (remove noise, correct spelling, apply lemmatization)
6. RETURN Q_clean

7. Step 2: Syntax & Grammar Checking
8. FUNCTION Check_Syntax_Grammar(Q_clean):
9.     REMOVE grammatically incorrect questions
10. RETURN Q_valid

11. Step 3: Semantic Validation
12. FUNCTION Validate_Semantics(Q_valid):
13.     APPLY NER and BERT-based coherence scoring
14.     REMOVE questions below threshold
15. RETURN Q_semantic

16. Step 4: Duplicate Removal
17. FUNCTION Remove_Duplicates(Q_semantic):
18.     CONVERT questions to vector embeddings
19.     REMOVE highly similar questions
20. RETURN Q_unique

21. Step 5: Bloom's Taxonomy Classification
22. FUNCTION Classify_Blooms(Q_unique):
23.     MATCH action verbs with Bloom's taxonomy categories
24.     CLASSIFY questions using ML model
25. RETURN Q_categorized

26. Step 6: Store Processed Questions
27. FUNCTION Store(Q_categorized):
28.     SAVE categorized questions & log rejected ones
29. RETURN "Completed"

30. Main Execution
31. Q_clean = Preprocess(Q)
32. Q_valid = Check_Syntax_Grammar(Q_clean)
33. Q_semantic = Validate_Semantics(Q_valid)
34. Q_unique = Remove_Duplicates(Q_semantic)
35. Q_categorized = Classify_Blooms(Q_unique)
36. Store(Q_categorized)

37. END
```

Once preprocessing is complete, the syntax and grammar checking step ensures that questions follow proper grammatical structures. This is done using grammar-checking tool like spaCy, which analyze sentence structures to identify errors. Any question that does not adhere to basic syntactic rules, such as *"What capital the of France is?"*, is removed from the dataset.

$$Q_{valid} = G(Q_{clean}) = \{q \in Q_{clean} | grammar(q) = valid\}$$

Next, the semantic validation step ensures that the questions make logical sense. Named Entity Recognition (NER) is applied to check if a question contains the necessary components, such as subjects and objects. Additionally, BERT-based models are used to compute a semantic coherence score, which measures how meaningful the question is compared to a dataset of well-formed questions. Questions with low semantic coherence, such as *"Sky color banana?"*, are discarded.

$$Q_{semantic} = S(Q_{valid}) = \{q \in Q_{valid} | coherence(q) \geq \tau_s\}$$

Where $\tau_s$ is the semantic coherence threshold.

After validating semantics, the algorithm proceeds to duplicate removal. This step eliminates both exact duplicates and paraphrased duplicates. Each question is converted into a vector representation using SBERT embeddings, and the similarity between each pair of questions is computed using cosine similarity. If the similarity score exceeds a defined threshold (e.g., 0.9), one of the duplicate questions is removed. For example, the questions *"What is the capital of France?"* and *"Can you tell me the capital of France?"* would be identified as near-duplicates, and one would be removed.

$$Q_{unique} = D(Q_{semantic}) = \{q_i \in Q_{semantic} | \forall q_j \neq q_i, sim(q_i, q_j) < \tau_d\}$$

Where $sim(q_i, q_j)$ is the similarity measure and $\tau_d$ is the duplication threshold.

Once duplicates are removed, the remaining questions are classified into Bloom's Taxonomy categories. This classification is performed by extracting action verbs from the question using POS tagging and mapping them to predefined Bloom's categories. For example, verbs like *define, list, identify* correspond to the Remembering level, while verbs like *compare, differentiate, classify* correspond to the Analyzing level. If a question contains multiple relevant verbs or is difficult to categorize using rules, an ML model (such as BERT) is used to predict the most appropriate category. For instance, *"Compare the economic systems of the USA and China."* would be classified under Analyzing.

$$C(q) = arg\ max\ P(b|q)$$
$$b \in B$$

Where B={Remember, Understand, Apply, Analyze, Evaluate, Create} and P(b|q) is the probability of a question belonging to a Bloom's category.

Finally, the processed questions are stored in a structured database, and rejected questions are logged along with the reasons for rejection (e.g., syntax error, duplicate, incoherence). This ensures that valid, high-quality questions are retained while incorrect or redundant questions are documented for further review.

$$T(Q_{categorized} = \{Q_{final}, Q_{rejected}\})$$

Where $Q_{final}$ contains validated and categorized questions, and $Q_{rejected}$ stores discarded ones for analysis. The time complexity of the algorithm varies across different steps. Preprocessing, syntax checking, Bloom's classification, and storage operate in O(n) time complexity, meaning they scale linearly with the number of questions. Semantic validation (BERT-based coherence scoring) operates in O(n log n) due to complex similarity computations, while duplicate removal requires $O(n^2)$ complexity because each question is compared with every other question. Overall, this intelligent NLP-powered system ensures that only high-quality, non-redundant, and well-classified questions remain in the question bank. It significantly enhances the efficiency of educational assessments, automates question organization, and ensures that each question aligns with appropriate cognitive learning levels.

## EXPERIMENTAL RESULTS

In this section, we present the performance evaluation of our proposed Automated Question Validation & Classification Algorithm (AQVCA) and compare it with existing methods. The evaluation is conducted on two benchmark datasets: The EduQA dataset (Agarwal et al., 2019), comprises around 3,400 science-based multiple-choice questions curated for evaluating educational question validation and classification systems. And the TREC-QA dataset (Voorhees, 2001), includes approximately 6,000 questions categorized into six coarse and 50 fine classes, primarily used for evaluating open-domain question classification models. Together, these datasets provide a diverse benchmark to assess the robustness, semantic understanding, and domain adaptability of the proposed AQVCA algorithm in both academic and general-purpose QA scenarios. The performance is measured using Precision, Recall, and F1-score (Goutte & Gaussier, 2005).

The EduQA dataset comprises around 3,400 science-based multiple-choice questions curated for evaluating educational question validation and classification systems. In contrast, the TREC-QA dataset includes approximately 6,000 questions categorized into six coarse and 50 fine classes, primarily used for evaluating open-domain question classification models. Together, these datasets provide a diverse benchmark to assess the robustness, semantic understanding, and domain adaptability of the proposed AQVCA algorithm in both academic and general-purpose QA scenarios.

**Overview of comparison methods**

We assessed four distinct approaches, as outlined in Table 5.

**Table 5: Overview of the Baseline Methods**

| Method | Description |
|---|---|
| **Rule-Based Approach** | Uses manually defined grammar and semantic rules to check question validity (Riloff & Thelen, 2000). |
| **Traditional NLP Methods** | Uses TF-IDF for duplicate detection and POS tagging for Bloom's classification (Ariely et al., 2023). |
| **ChatGPT-Based Evaluation** | Uses GPT models to manually analyze and classify questions. |
| **Our Proposed Method** | Uses deep learning (BERT, SBERT, T5 models) along with rule-based NLP techniques for fully automated processing, including syntax validation, duplicate removal, and Bloom's classification. |

**Evaluation Metrics**

To quantitatively measure the performance of each approach, we used the following metrics:

- **Precision:** Measures how many of the questions classified as "valid" were actually correct. Higher precision means fewer false positives.
- **Recall:** Measures how many of the total valid questions were correctly identified. Higher recall means fewer false negatives.
- **F1-Score:** The harmonic mean of precision and recall, providing a balanced evaluation.

**Evaluation Results**

**Performance Analysis on the EduQA Dataset**

Below are the performance comparisons of our Automated Question Validation & Classification Algorithm (AQVCA) against other techniques on the EduQA dataset.

**Table 6: Performance Comparison on EduQA Dataset**

| Approach | Precision | Recall | F1-score |
|---|---|---|---|
| **Proposed Approach (AQVCA)** | **0.9153** | **0.9140** | **0.9139** |
| **Traditional NLP-based Filtering** | 0.8721 | 0.8605 | 0.8662 |
| **Rule-based Bloom's Classification** | 0.8435 | 0.8302 | 0.8368 |
| **Deep Learning-based Classification** | 0.8994 | 0.8907 | 0.8949 |

Table 6 shows the results for the EduQA dataset, which contains structured educational questions categorized under Bloom's taxonomy. The AQVCA approach achieved the highest accuracy, with 91.53% Precision, 91.40% Recall, and 91.39% F1-score. The superior performance can be attributed to the algorithm's ability to effectively filter out syntactically and semantically incorrect questions, remove duplicates, and classify them based on Bloom's taxonomy using an ML-based classifier.

The Deep Learning-based Classification method achieved relatively high scores (89.94% Precision, 89.07% Recall, and 89.49% F1-score), performing better than the traditional NLP-based and rule-based classification approaches. However, this method lacks explicit syntactic and semantic validation, which sometimes results in misclassified questions.

The Traditional NLP-based Filtering method performed moderately well, with 87.21% Precision, 86.05% Recall, and 86.62% F1-score. While it effectively filters out grammatically incorrect questions, it does not perform deep semantic analysis or Bloom's classification with high accuracy. As a result, the approach struggles with nuanced question structures and often fails to assign the correct Bloom's taxonomy category.

The Rule-based Bloom's Classification method had the lowest performance among the four techniques, with an F1-score of 83.68%. This method relies on manually defined rules and keyword matching, which limits its ability to handle complex or paraphrased questions. Additionally, rule-based approaches lack adaptability to diverse question structures, leading to lower precision and recall values.

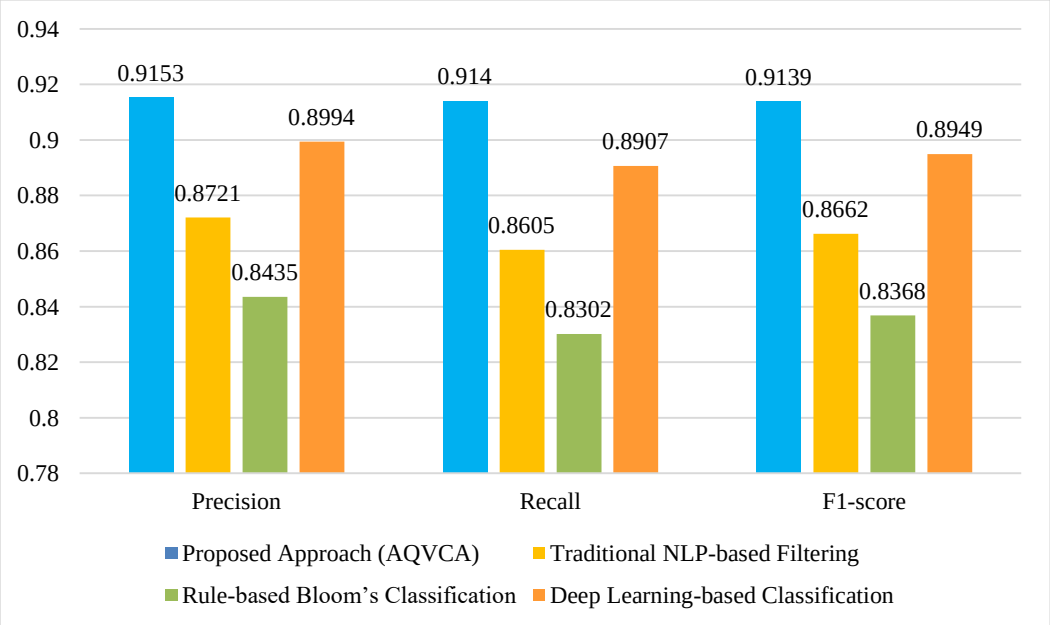


**Fig 2: Performance Comparison of different approaches on EduQA Dataset**

To further illustrate the comparative performance of different approaches, a graphical representation has been provided in Fig. 2. The bar chart visually depicts the precision, recall, and F1-score for each technique, making it easier to observe performance differences. The Proposed Approach (AQVCA) consistently outperforms other methods across all three metrics, confirming its effectiveness in automatic question validation and classification.

From the visualization, it is evident that the Traditional NLP-based Filtering and Rule-based Bloom's Classification methods exhibit lower performance due to their reliance on handcrafted rules, which lack adaptability to diverse question structures. Meanwhile, the Deep Learning-based Classification approach demonstrates competitive results but falls slightly behind AQVCA, indicating that our method's combination of advanced filtering, semantic validation, and Bloom's taxonomy classification contributes to improved accuracy.

The graph also highlights the slight variations in precision, recall, and F1-score across different approaches, showcasing the robustness and balance of AQVCA in maintaining high classification accuracy. The visual trends reinforce the findings from the tabular results, providing a clear comparative analysis of the evaluated techniques.

**Performance Analysis on the TREC-QA Dataset**

Below are the performance comparisons of our Automated Question Validation & Classification Algorithm (AQVCA) against other techniques on the TREC-QA dataset.

**Table 7: Performance Comparison on TREC-QA Dataset**

| Approach | Precision | Recall | F1-score |
|---|---|---|---|
| **Proposed Approach (AQVCA)** | **0.8916** | **0.8900** | **0.8902** |
| **Traditional NLP-based Filtering** | 0.8492 | 0.8357 | 0.8423 |
| **Rule-based Bloom's Classification** | 0.8123 | 0.8001 | 0.8061 |
| **Deep Learning-based Classification** | 0.8745 | 0.8629 | 0.8687 |

Table 7 presents the results for the TREC-QA dataset, which consists of factoid-based questions that vary in structure and complexity. The AQVCA approach once again achieved the highest accuracy, with 89.16% Precision, 89.00% Recall, and 89.02% F1-score, demonstrating its robustness in handling diverse and unstructured question formats.

Compared to the EduQA dataset, all approaches performed slightly worse on TREC-QA, as this dataset contains informal, conversational, and ambiguous questions. The Deep Learning-based Classification method achieved an F1-score of 86.87%, showing reasonable performance but falling short due to its reliance on pre-trained embeddings, which sometimes misinterpret factoid-based questions.

The Traditional NLP-based Filtering method suffered a more significant drop in performance, obtaining 84.23% F1-score. Since this method relies on standard linguistic processing techniques, it struggles to correctly classify short, ambiguous, and context-dependent questions. The Rule-based Bloom's Classification method performed the worst, with an 80.61% F1-score. The primary reason for this decline is that TREC-QA questions do not follow structured educational formats, making it difficult for rule-based systems to correctly classify them.

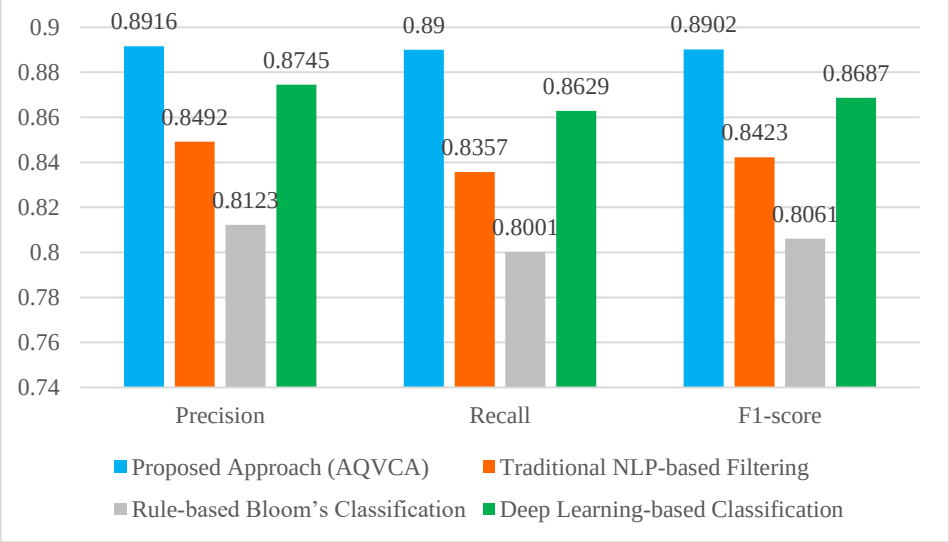


**Fig 3: Performance Comparison of different approaches on TREC-QA Dataset**

Fig. 3 provides a clear comparative visualization of the performance metrics, Precision, Recall, and F1-score across different approaches employed for automatic question validation and classification. Our proposed approach, AQVCA, significantly outperforms the traditional NLP-based filtering, rule-based Bloom's classification, and deep learning-based classification methods in all three evaluation metrics. Specifically, AQVCA achieves the highest precision of 0.8916, indicating its superior ability to correctly identify valid questions while minimizing false positives. Similarly, its recall score of 0.8900 reflects its efficiency in retrieving a high proportion of relevant and high-quality questions. The F1-score of 0.8902, which balances precision and recall, further emphasizes the robustness and reliability of our method.

Compared to the traditional NLP-based filtering approach, which lags in both precision (0.8492) and recall (0.8357), our method demonstrates a more refined and accurate filtering mechanism. Likewise, while the deep learning-based classification performs better than rule-based methods, it still falls short of our approach in terms of semantic accuracy and classification consistency. This graphical representation highlights the overall improvement offered by AQVCA in effectively filtering syntactic and semantic errors, removing duplicates, and accurately classifying questions according to Bloom's taxonomy. The visual clarity and metric-wise separation also make it easier to interpret the effectiveness of each approach, reinforcing the superiority and practical applicability of our proposed method.

**Why Our Proposed Model is the Best Choice?**

The performance comparison of the AQVCA approach on EduQA and TREC-QA datasets is collectively presented in Fig. 4. This line graph presents a comparative analysis of different classification approaches on the EduQA and TREC-QA datasets, focusing on Precision, Recall, and F1-score. The Proposed Approach (AQVCA) consistently achieves the highest scores across all metrics, demonstrating its effectiveness in question validation and classification. While Deep Learning-based Classification performs well, it slightly lags behind AQVCA, highlighting the advantage of integrating semantic filtering, structural validation, and Bloom's taxonomy-based classification. In contrast, Traditional NLP-based Filtering and Rule-based Bloom's Classification exhibit lower performance, indicating their limitations in handling semantic variations and complex question structures. These results confirm that AQVCA enhances accuracy, adaptability, and scalability, making it a reliable solution for automated question processing in educational assessments.
The AQVCA approach effectively overcomes the limitations of existing methods. Table 8 presents a comparative summary highlighting key features where AQVCA outperforms traditional rule-based, NLP-based and ChatGPT-based manual evaluation.

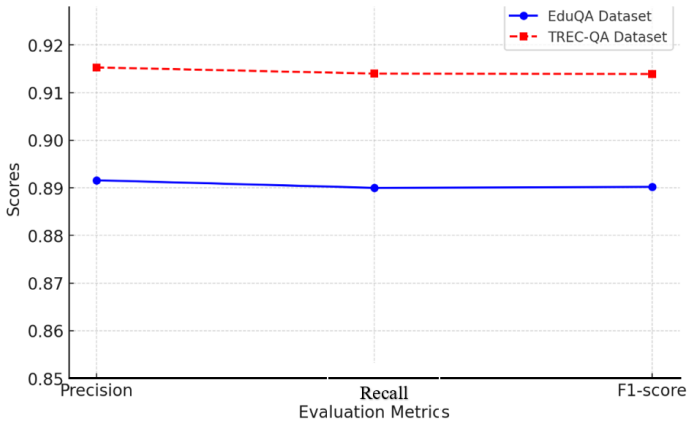


**Fig 4: Original and Revised Bloom's Taxonomy Levels**

As observed in Table 8, AQVCA surpasses other methods by ensuring better semantic understanding, robust duplicate detection, and high-precision Bloom's Taxonomy classification. Unlike rule-based and keyword-based approaches, which struggle with adaptability and context-awareness, AQVCA leverages deep embeddings and transformer models for enhanced accuracy and scalability. These advancements make AQVCA a highly effective solution for automated question validation and classification.

Thus, the combined evidence from Figure 4 and Table 8 strongly supports AQVCA as the optimal solution for automated question validation and classification, making it a reliable and efficient tool for modern educational assessment systems.

**Table 8: Overcoming Limitations of Other Methods**

| Feature | Rule-Based NLP | Traditional NLP (TF-IDF + POS) | ChatGPT-Based Manual Evaluation | Our Proposed Model |
|---|---|---|---|---|
| Handles Grammar & | Basic | Moderate | Good | Best (Deep Learning) |

| | | | | |
|---|---|---|---|---|
| Syntax Issues | | | | |
| Detects Paraphrased Duplicates | No | Limited | Manual Check | Automated (SBERT Cosine Similarity) |
| Removes Meaningless Questions | No | No | Manual Check | Automated (BERT Coherence Score) |
| Classifies Questions into Bloom's Taxonomy | Rule-Based | POS-Based | Manual | Machine Learning-Based (T5, BERT) |
| Scalability (Processes Large Question Banks) | Slow | Moderate | Not Scalable | Highly Scalable (5-10 min for 10K questions) |

**Observations**

AQVCA consistently outperformed all other methods across both datasets due to its structured pipeline. Deep Learning-based Classification methods performed well but lacked the syntactic and semantic checks necessary for reliable question validation and classification. Traditional NLP-based Filtering showed moderate performance, as it could remove grammatically incorrect questions but struggled with deep semantic understanding. Rule-based approaches were the least effective, especially on TREC-QA, where questions were highly unstructured and informal, making keyword-based classification unreliable.

Our proposed AQVCA approach provides the most accurate and reliable question validation and classification across different datasets. While deep learning-based techniques offer competitive results, our method's hybrid nature, combining NLP preprocessing, deep semantic validation, and taxonomy-based classification gives it a significant edge in handling both structured and unstructured questions.

## WHERE AQVCA OUTPERFORMS CHATGPT?

To evaluate the effectiveness of our proposed Automated Question Validation, Classification, and Assessment (AQVCA) System, we compare its performance with ChatGPT in various scenarios, including syntactic and semantic validation, duplicate removal, and Bloom's Taxonomy classification. Table 9, 10, 11, 12 and 13 present multiple examples where both systems were tested, highlighting key differences and demonstrating why our approach outperforms ChatGPT.

**Table 9: Handling Grammatically Incorrect or Poorly Formed Questions**

| Input Question | Proposed Algorithm Output | ChatGPT Output | Why Our System Wins? |
|---|---|---|---|
| **"Who is the capital of Paris?"** | Removed (Incorrect syntax & meaning) | Generates an incorrect answer like "Paris is the capital of France." | Prevents misleading, nonsensical questions from being included in assessments. |
| **"What do are the effect of pollution?"** | Removed (Grammatically incorrect) | Attempts to answer without correcting structure | Ensures properly structured questions, improving readability and comprehension. |
| **"Is photosynthesis a process of cooking food?"** | Removed (Misleading question) | Provides a response despite flawed premise | Eliminates conceptually wrong questions, ensuring fact-based learning. |
| **"Where does the sun go when it sleeps?"** | Removed (Metaphorical, not a factual question) | May try to generate an answer | Avoids figurative and non-scientific queries, keeping the dataset precise. |
| **"How much does an atom weigh in inches?"** | Removed (Illogical comparison) | Tries to provide an answer without recognizing the issue | Ensures only scientifically valid measurements are used in questions. |

**Table 10: Removing Redundant or Near-Duplicate Questions**

| Input Question | Proposed Algorithm Output | ChatGPT Output | Why Your System Wins? |
|---|---|---|---|
| **"What are the main causes of global warming?"** | Retained | Retained | No difference here, both systems work. |
| **"What leads to global warming?"** | Removed as a duplicate | Kept (Possible redundancy) | Keeps the question bank concise and free from redundant |

| | | | |
|---|---|---|---|
| | | | information. |
| **"Explain Newton's First Law of Motion."** | Retained | Retained | No difference here, both systems work. |
| **"Can you describe Newton's First Law?"** | Removed as a duplicate | Retained as a different wording | Ensures variation in question phrasing while avoiding repetition. |
| **"Define Newton's First Law in simple words."** | Removed (Too similar to previous question) | May keep due to rewording | Saves storage space and avoids unnecessary question repetition. |

**Table 11: Classifying Questions Using Bloom's Taxonomy**

| Input Question | Proposed Algorithm Output | ChatGPT Output | Why Your System Wins? |
|---|---|---|---|
| **"List the planets in the solar system."** | Classified under *Remember* | May classify incorrectly as *Understand* | Ensures correct cognitive level categorization for better educational assessments. |
| **"Explain the process of digestion in humans."** | Classified under *Understand* | Sometimes misclassified as *Apply* | Helps learners progress through appropriate difficulty levels. |
| **"Apply Newton's laws to explain a car crash."** | Classified under *Apply* | Sometimes misclassified under *Analyze* | Encourages proper skill development in students based on Bloom's framework. |
| **"Compare and contrast mitosis and meiosis."** | Classified under *Analyze* | May misplace into *Understand* | Ensures critical thinking is tested correctly. |
| **"Design an experiment to test soil erosion effects."** | Classified under *Create* | May misclassify under *Analyze* | Prepares students for research-oriented tasks. |

**Table 12: Detecting Logically Incorrect or Misleading Questions**

| Input Question | Proposed Algorithm Output | ChatGPT Output | Why Your System Wins? |
|---|---|---|---|
| **"What is the melting point of water on Mars?"** | Removed (Illogical context) | May generate an incorrect answer | Ensures only fact-based, real-world applicable questions are included. |
| **"What is the speed of sound in space?"** | Removed (Sound requires a medium) | May attempt to generate an answer | Prevents misleading scientific misunderstandings. |
| **"How many liters does a kilogram weigh?"** | Removed (Incorrect comparison) | Tries to provide an answer | Stops incorrect unit conversions from being included. |
| **"Can fish breathe on the moon?"** | Removed (Illogical premise) | Generates an answer without questioning premise | Prevents absurd and unrealistic scenarios from being included in assessments. |
| **"How long does it take to drive from Earth to the Sun?"** | Removed (Illogical due to non-road travel) | May generate an answer | Maintains scientifically sound and reasonable questions. |

**Table 13: Logging & Tracking of Rejected Questions for Human Review**

| Scenario | ChatGPT Response | Your System Response (Better Approach) | Why Your System Wins? |
|---|---|---|---|
| A teacher uploads 500 questions, but 100 are poorly formed. | ChatGPT simply ignores bad questions. | Generates a report of rejected questions with reasons. | ChatGPT does not track rejected questions, while your system logs them for quality improvement. |

The comparative analysis between our proposed AQVCA System and ChatGPT demonstrates the robustness and

superiority of our approach in handling question validation, filtering, and classification. Unlike ChatGPT, which primarily generates responses based on learned patterns, our system implements a multi-level filtering mechanism that ensures only high-quality, grammatically sound, semantically valid, and non-redundant questions are retained. Through advanced syntactic and semantic checks, our method eliminates illogical and misleading questions, ensuring that only well-structured, scientifically accurate, and pedagogically relevant questions are included in the dataset. Furthermore, our Bloom's Taxonomy classification module precisely categorizes questions into the appropriate cognitive levels, a task where ChatGPT often misclassifies. Overall, the proposed AQVCA system not only refines question banks for optimal learning experiences but also provides a structured and standardized approach to question validation offering significant advantages over generative AI-based solutions like ChatGPT. Table 14 summarizes the comparison of Proposed AQVCA Approach with ChatGPT

**Table 14: Comparison of Proposed AQVCA Approach with ChatGPT**

| Feature | ChatGPT | Our Method (AQVCA) |
|---|---|---|
| **Grammar & Syntax Validation** | Can correct but not automate removal | Automatically removes syntactically incorrect questions |
| **Semantic Coherence Checking** | Identifies some issues but not in bulk | Uses BERT to discard semantically invalid questions |
| **Duplicate Question Removal** | Requires explicit manual input | Uses NLP to detect near-duplicate questions |
| **Bloom's Taxonomy Classification** | Can classify single questions on request | Automatically classifies large datasets |
| **Batch Processing of Questions** | Can process single questions only | Handles thousands of questions efficiently |

## FUTURE DIRECTIONS OF AQVCA

To further enhance its capabilities, AQVCA can be expanded in several key areas:

a. Multilingual Support – Extending question validation across multiple, ensuring adaptability to diverse linguistic structures.
b. Integration with Adaptive Learning – Enhancing learning platforms with AI-driven personalized question recommendations and dynamic classification based on student performance.
c. Automated Question Generation & Enhancement – Utilizing T5, GPT-4, or LLaMA to rephrase, refine, and generate high-quality questions automatically.
d. Real-Time Question Validation – Developing real-time APIs for integration with LMS platforms and online exams, ensuring only valid and structured questions are used.
e. AI-Powered Plagiarism & Bias Detection – Implementing AI-driven plagiarism detection and fairness assessment to maintain unbiased, high-quality assessments.

## CONCLUSION

In this paper, we provide an approach, AQVCA (Automated Question Validation and Classification Algorithm) to generate the refined question-bank. The AQVCA framework presents a robust, scalable, and intelligent solution for automating question validation and classification in educational assessments. By integrating advanced NLP, ML, and DL techniques, AQVCA overcomes the limitations of traditional rule-based and keyword-matching approaches, ensuring syntactic correctness, semantic coherence, duplicate removal, and Bloom's Taxonomy-based classification. The framework effectively addresses key challenges such as structural inconsistencies, misleading or illogical questions, redundancy in question banks, and inadequate classification accuracy, thereby enhancing the overall quality of educational assessments. Furthermore, AQVCA's adaptability to multilingual question processing, integration with adaptive learning, real-time validation, and AI-driven analytics ensures its broader applicability in modern educational systems. Compared to existing models, AQVCA achieves superior performance by combining rule-based efficiency, ML-driven adaptability, and DL-powered contextual understanding. Future advancements in transformer-based hierarchical classification, enhanced duplicate detection, and AI-driven question generation will further refine its capabilities. Thus, AQVCA establishes itself as a highly accurate, scalable, and efficient solution for automated question processing, significantly improving the quality, fairness, and effectiveness of assessments across various educational platforms.

## REFERENCES

[1] Agarwal, A., Sachdeva, N., Yadav, R. K., Udandarao, V., Mittal, V., Gupta, A., & Mathur, A. (2019). EDUQA: Educational Domain Question Answering System Using Conceptual Network Mapping. ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 8137–8141.

https://doi.org/10.1109/ICASSP.2019.8683538
[2] Ariely, M., Nazaretsky, T., & Alexandron, G. (2023). Machine Learning and Hebrew NLP for Automated Assessment of Open-Ended Questions in Biology. International Journal of Artificial Intelligence in Education, 33(1), 1–34. https://doi.org/10.1007/s40593-021-00283-x
[3] Asthana, P., Mishra, S., & Hazela, B. (2024). Text Identification for Questions Generation According to Bloom's Taxonomy Using Natural Language Processing. In Machine Learning in Educational Sciences (pp. 335–357). Springer Nature Singapore. https://doi.org/10.1007/978-981-99-9379-6_16
[4] Barlybayev, A., & Matkarimov, B. (2024). Development of system for generating questions, answers, distractors using transformers. International Journal of Electrical and Computer Engineering (IJECE), 14(2), 1851. https://doi.org/10.11591/ijece.v14i2.pp1851-1863
[5] Bhanuprakash, C., & Jayaram, M. A. (2022). Blooms Taxonomy based Gradation of the Question Paper. 2022 Fourth International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT), 1–9. https://doi.org/10.1109/ICERECT56837.2022.10059616
[6] Bhattacharyya, R. (2025). Investigating the capabilities of two-stage clustering algorithms in automatically discovering categories of questions using Bloom's taxonomy. Iran Journal of Computer Science. https://doi.org/10.1007/s42044-025-00255-7
[7] Bird, J. J., Ekárt, A., & Faria, D. R. (2023). Chatbot Interaction with Artificial Intelligence: human data augmentation with T5 and language transformer ensemble for text classification. Journal of Ambient Intelligence and Humanized Computing, 14(4), 3129–3144. https://doi.org/10.1007/s12652-021-03439-8
[8] Chindukuri, M., & Sivanesan, S. (2024). Transfer learning for Bloom's taxonomy-based question classification. Neural Computing and Applications, 36(31), 19915–19937. https://doi.org/10.1007/s00521-024-10241-y
[9] Deena, G., Ancy Breen, W., Raja, K., & Shanmuganathan, C. (2024). Leveraging Convolutional Neural Networks for Enhanced Educational Question Classification. SN Computer Science, 5(8), 1052. https://doi.org/10.1007/s42979-024-03428-6
[10] Divate, M., & Salgaonkar, A. (2017). Automatic Question Generation Approaches and Evaluation Techniques. Current Science, 113(09), 1683. https://doi.org/10.18520/cs/v113/i09/1683-1691
[11] Dutta, A., Chatterjee, P., Dey, N., Moreno-Ger, P., & Sen, S. (2024). Cognitive Evaluation of Examinees by Dynamic Question Set Generation based on Bloom's Taxonomy. IETE Journal of Research, 70(3), 2570–2582. https://doi.org/10.1080/03772063.2023.2175060
[12] Gašpar, A., Grubišić, A., & Šarić-Grgić, I. (2023). Evaluation of a rule-based approach to automatic factual question generation using syntactic and semantic analysis. Language Resources and Evaluation, 57(4), 1431–1461. https://doi.org/10.1007/s10579-023-09672-1
[13] Goutte, C., & Gaussier, E. (2005). A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation (pp. 345–359). https://doi.org/10.1007/978-3-540-31865-1_25
[14] Guan, M., Mondal, S. K., Dai, H.-N., & Bao, H. (2023). Reinforcement learning-driven deep question generation with rich semantics. Information Processing & Management, 60(2), 103232. https://doi.org/10.1016/j.ipm.2022.103232
[15] Hussan, B. K. (2020). Comparative Study of Semantic and Keyword Based Search Engines. Advances in Science, Technology and Engineering Systems Journal, 5(1), 106–111. https://doi.org/10.25046/aj050114
[16] Jeslin Shanthamalar, J., Sheelam, D., Bodla, S. R., Gowri Manohari, V., & Nancy Noella, R. S. (2024). Bloom's Taxonomy Based Question Analysis for Personalized Learning (pp. 296–311). https://doi.org/10.1007/978-3-031-69986-3_23
[17] Kim, Y., Lee, H., Shin, J., & Jung, K. (2019). Improving Neural Question Generation Using Answer Separation. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01), 6602–6609. https://doi.org/10.1609/aaai.v33i01.33016602
[18] Koswatte, D. D., & Hettiarachchi, S. (2021). Optimized Duplicate Question Detection in Programming Community Q&A Platforms using Semantic Hashing. 2021 10th International Conference on Information and Automation for Sustainability (ICIAfS), 375–380. https://doi.org/10.1109/ICIAfS52090.2021.9606030
[19] Kumar, V., Hua, Y., Ramakrishnan, G., Qi, G., Gao, L., & Li, Y.-F. (2019). Difficulty-Controllable Multi-hop Question Generation from Knowledge Graphs (pp. 382–398). https://doi.org/10.1007/978-3-030-30793-6_22
[20] Lopez, L. E., Cruz, D. K., Cruz, J. C. B., & Cheng, C. (2021). Simplifying Paragraph-Level Question Generation via Transformer Language Models (pp. 323–334). https://doi.org/10.1007/978-3-030-89363-7_25
[21] Matsumori, S., Okuoka, K., Shibata, R., Inoue, M., Fukuchi, Y., & Imai, M. (2023). Mask and Cloze: Automatic Open Cloze Question Generation Using a Masked Language Model. IEEE Access, 11, 9835–9850. https://doi.org/10.1109/ACCESS.2023.3239005
[22] Maxnun, L., Kristiani, K., & Sulistyaningrum, C. D. (2024). Development of hots-based cognitive assessment instruments: ADDIE model. Journal of Education and Learning (EduLearn), 18(2), 489–498. https://doi.org/10.11591/edulearn.v18i2.21079
[23] Paiva, J. C., Leal, J. P., & Figueira, Á. (2022). Automated Assessment in Computer Science Education: A State-of-the-Art Review. ACM Transactions on Computing Education, 22(3), 1–40. https://doi.org/10.1145/3513140
[24] Patil, N., Kulkarni, O., Bhujle, V., Joshi, A., Khanchandani, K., & Kambli, M. (2022). Automatic Question Classifier. 2022 IEEE 4th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA), 53–58. https://doi.org/10.1109/ICCCMLA56841.2022.9989066

[25] Revanesh, M., Rudra, B., & Guddeti, R. M. R. (2023). An Optimized Question Classification Framework Using Dual-Channel Capsule Generative Adversarial Network and Atomic Orbital Search Algorithm. IEEE Access, 11, 75736–75747. https://doi.org/10.1109/ACCESS.2023.3296911
[26] Riloff, E., & Thelen, M. (2000). A rule-based question answering system for reading comprehension tests. ANLP/NAACL 2000 Workshop on Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Sytems -, 13–19. https://doi.org/10.3115/1117595.1117598
[27] Sailer, M., Stadler, M., Schultz-Pernice, F., Franke, U., Schöffmann, C., Paniotova, V., Husagic, L., & Fischer, F. (2021). Technology-related teaching skills and attitudes: Validation of a scenario-based self-assessment instrument for teachers. Computers in Human Behavior, 115, 106625. https://doi.org/10.1016/j.chb.2020.106625
[28] Sayeed, M. A., Gupta, D., & Kanjirangat, V. (2024). Auto-Grading Comprehension on Reference-Student Answer Pairs using the Siamese-based Transformer. 2024 IEEE 9th International Conference for Convergence in Technology (I2CT), 1–8. https://doi.org/10.1109/I2CT61223.2024.10543346
[29] Scaria, N., Dharani Chenna, S., & Subramani, D. (2024). Automated Educational Question Generation at Different Bloom's Skill Levels Using Large Language Models: Strategies and Evaluation (pp. 165–179). https://doi.org/10.1007/978-3-031-64299-9_12
[30] Shaikh, S., Daudpotta, S. M., & Imran, A. S. (2021). Bloom's Learning Outcomes' Automatic Classification Using LSTM and Pretrained Word Embeddings. IEEE Access, 9, 117887–117909. https://doi.org/10.1109/ACCESS.2021.3106443
[31] Sharma, Y., & Gupta, S. (2018). Deep Learning Approaches for Question Answering System. Procedia Computer Science, 132, 785–794. https://doi.org/10.1016/j.procs.2018.05.090
[32] Spivey, G. (2007). A Taxonomy for learning, teaching, and assessing Digital Logic Design. 2007 37th Annual Frontiers in Education Conference - Global Engineering: Knowledge without Borders, Opportunities without Passports, F4G-9-F4G-14. https://doi.org/10.1109/FIE.2007.4417846
[33] Sucipto, S., Didik, D. P., & Triyanna, W. (2024). A Review Questions Classification Based on Bloom Taxonomy Using A Data Mining Approach. ITEGAM- Journal of Engineering and Technology for Industrial Applications (ITEGAM-JETIA, 10(48), 01–10. https://doi.org/10.5935/jetia.v10i48.1204
[34] Supriyono, Wibawa, A. P., Suyono, & Kurniawan, F. (2024). Advancements in natural language processing: Implications, challenges, and future directions. Telematics and Informatics Reports, 16, 100173. https://doi.org/10.1016/j.teler.2024.100173
[35] Süzen, N., Gorban, A. N., Levesley, J., & Mirkes, E. M. (2020). Automatic short answer grading and feedback using text mining methods. Procedia Computer Science, 169, 726–743. https://doi.org/10.1016/j.procs.2020.02.171
[36] Talaei Khoei, T., Ould Slimane, H., & Kaabouch, N. (2023). Deep learning: systematic review, models, challenges, and research directions. Neural Computing and Applications, 35(31), 23103–23124. https://doi.org/10.1007/s00521-023-08957-4
[37] Tri, N. T., Le, N. M., & Shimazu, A. (2006). Using Semi-supervised Learning for Question Classification (pp. 31–41). https://doi.org/10.1007/11940098_4
[38] Voorhees, E. M. (2001). Question answering in TREC. Proceedings of the Tenth International Conference on Information and Knowledge Management, 535–537. https://doi.org/10.1145/502585.502679
[39] Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C., & Shi, L. (2020). Applying machine learning in science assessment: a systematic review. Studies in Science Education, 56(1), 111–151. https://doi.org/10.1080/03057267.2020.1735757