Original Article
ISSN (Online): 2582-7472

ShodhKosh: Journal of Visual and Performing Arts
May 2024 5(5), 1687–1697

AUTOMATED BLOOM'S TAXONOMY-BASED QUESTION GENERATION FOR COURSE OUTCOME ATTAINMENT IN OBE FRAMEWORKS

Blessy Paul P [1], Cini Kurian [2], John T Abraham [3]

[1] School of Computer Sciences, Mahatma Gandhi University, Kottayam, Kerala, India
[2] AL-Ameen College, Edathala, Aluva, Ernakulam, Kerala, India
[3] Bharata Mata College, Thrikkakara, Ernakulam, Kerala, India

ABSTRACT

The creation of assessment questions that align with Bloom's taxonomy levels and achieve Course Outcomes (COs) is a critical yet complex task in Outcome-Based Education (OBE). Traditional manual methods, reliant on subject experts, are time-consuming and prone to gaps in addressing all COs or Bloom's levels. While Large Language Models (LLMs) like ChatGPT can generate questions, they lack access to private data, including prescribed textbooks and syllabi, potentially leading to questions beyond the scope of the curriculum. This paper presents a novel system leveraging Retrieval-Augmented Generation (RAG) to automate the generation of Bloom's taxonomy-based questions within the syllabus scope, ensuring comprehensive CO attainment. The proposed system integrates a vector database to store private data, including scanned textbooks, syllabi, Bloom's taxonomy levels, and COs. The RAG model, trained on this curated dataset, generates questions that fulfill the cognitive, psychomotor, and affective domain requirements specified in the syllabus. This approach not only ensures alignment with educational objectives but also significantly reduces the manual effort involved in question preparation. The system's efficacy is demonstrated through its ability to produce high-quality, targeted questions that effectively support OBE evaluation and enhance educational quality. This innovation addresses a critical gap in automated question generation for modern education systems.

Keywords: Question Generation, Outcome-Based Education (OBE), Course Outcomes (COS), Retrieval-Augmented Generation (RAG) Model, Large Language Model (LLM)

Corresponding Author

Blessy Paul P, blessypaul91@gmail.com

DOI
10.29121/shodhkosh.v5.i5.2024.6108

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

1. INTRODUCTION

In the realm of education, particularly within Outcome-Based Education (OBE), the formulation of assessment questions that align with Bloom's Taxonomy and ensure the attainment of Course Outcomes (COs) is paramount. Bloom's Taxonomy provides a hierarchical classification of cognitive skills, ranging from basic knowledge recall to complex analytical abilities, serving as a foundational framework for educators to structure learning objectives and assessments. Traditionally, the development of such assessment tools has been a manual endeavor, heavily reliant on subject matter experts meticulously crafting questions that correspond to specific cognitive levels and course outcomes. This manual process, while thorough, is inherently time-consuming and susceptible to human error. Educators may inadvertently overlook certain COs or fail to address all levels of Bloom's Taxonomy, leading to assessments that do not fully encapsulate the intended learning objectives. The increasing complexity of curricula and the diverse cognitive skills they

aim to develop further exacerbate these challenges, underscoring the need for more efficient and comprehensive methods of question generation.

In recent years, advancements in artificial intelligence, particularly the emergence of Large Language Models (LLMs) like ChatGPT, have introduced new possibilities for automating the question generation process. LLMs are capable of generating human-like text based on a given input, offering a potential solution to the labor-intensive task of question creation. However, these models operate primarily on publicly available data and lack access to proprietary educational materials such as prescribed textbooks and specific course syllabi. Consequently, the questions generated by LLMs may deviate from the precise scope of a course, potentially misaligning with the targeted COs and cognitive levels.

To address these limitations, this paper proposes an innovative system that leverages Retrieval-Augmented Generation (RAG) to automate the creation of assessment questions. RAG combines the generative capabilities of LLMs with a retrieval mechanism that accesses a curated database of private educational content, including scanned textbooks, detailed syllabi, Bloom's Taxonomy classifications, and defined course outcomes. By integrating these resources, the system can generate questions that are not only contextually relevant but also precisely aligned with the educational objectives of a course. The core component of this system is a vector database that stores the private educational materials in a structured format conducive to efficient retrieval. When tasked with generating a question, the RAG model queries this database to extract pertinent information, which it then uses to inform and ground the generated content. This approach ensures that the questions adhere strictly to the prescribed syllabus and accurately reflect the desired cognitive levels as outlined in Bloom's Taxonomy.

Recent studies have explored various methodologies for automated question generation within educational contexts. For instance, Gnanasekaran et al. (2021) developed an Automatic Question Generation (AQG) system utilizing a rule-based approach aligned with Bloom's Taxonomy, demonstrating the potential for automation in educational assessments. Similarly, Henkel et al. (2023) investigated the application of retrieval-augmented generation to enhance math question-answering, highlighting the trade-offs between groundedness and human preference in educational content generation. Additionally, Dhainje et al. (2024) proposed an automated question paper generation system incorporating weighting based on Bloom's Taxonomy, emphasizing the importance of aligning questions with cognitive levels to promote deeper understanding and critical thinking. Building upon these foundational works, the system presented in this paper aims to enhance the automation of question generation by ensuring comprehensive coverage of all COs and Bloom's Taxonomy levels, thereby supporting a more robust and objective assessment framework within OBE. By reducing the manual effort required and minimizing the potential for human error, this approach seeks to improve the quality and effectiveness of educational assessments, ultimately contributing to better learning outcomes.

2. LITERATURE REVIEW

The automation of exam question generation has been a focus of extensive research, driven by the need to streamline assessment creation while maintaining alignment with educational objectives. Early approaches, such as template-driven and genetic algorithm-based systems, provided structured question generation but were limited in flexibility and adaptability to diverse cognitive and course-specific requirements. Advances in AI brought semantic and contextual improvements through models like BERT and joint question-answering frameworks, yet these systems often relied on publicly available data, leading to misalignment with prescribed syllabi. Recent efforts to integrate Bloom's Taxonomy and retrieval-augmented generation have shown promise but fall short in dynamically incorporating proprietary educational materials. These limitations underscore the need for a system capable of addressing cognitive complexity and course outcome alignment while leveraging private educational data, a gap that the proposed RAG-based system seeks to fill.

In recent years, the automation of exam question generation has gained significant attention due to its potential to enhance the efficiency and precision of educational assessments. Hussein et al. (2014) developed an Automatic English Question Generation System using a template-driven approach, ensuring grammatical accuracy and syntactic structure. However, the rigidity of templates limited the diversity and adaptability of questions, failing to address the cognitive complexity levels outlined in Bloom's Taxonomy. Similarly, Abd Rahim et al. (2017) employed a genetic algorithm for question generation, optimizing the selection of questions based on predefined criteria. While effective for certain contexts, this approach lacked the flexibility to align with specific course outcomes and taxonomy levels, presenting a gap in meeting diverse educational needs.

AI-based models have further advanced the field, with Wang et al. (2017) proposing a joint model for question answering and generation that leveraged shared representations for improved contextual relevance. Although innovative, this model did not incorporate course-specific constraints, limiting its applicability in educational contexts. Chan and Fan (2019) introduced a recurrent BERT-based model that demonstrated enhanced semantic coherence in question generation. Despite this progress, reliance on publicly available datasets led to questions misaligned with prescribed syllabi, a recurring challenge in AI-driven systems.

Kurdi et al. (2020) conducted a comprehensive review of automatic question generation for educational purposes, identifying persistent challenges, including the lack of systems capable of addressing both cognitive complexity and course-specific outcomes. More recent efforts have attempted to integrate Bloom's Taxonomy into question generation. For instance, Mohandas et al. (2024) introduced a system incorporating weighting based on taxonomy levels, enhancing alignment with cognitive objectives. However, this system relied on static datasets and lacked the adaptability to handle syllabus updates dynamically. Levonian et al. (2023) explored retrieval-augmented generation (RAG) for grounded question generation, demonstrating its potential to contextualize outputs within specific knowledge domains. Nevertheless, their work primarily addressed general knowledge tasks, leaving educational datasets and proprietary content underexplored.

Despite these advancements, existing approaches exhibit several limitations, including rigid frameworks, dependence on public data, and insufficient alignment with cognitive complexity, and inadequate integration of proprietary educational content. The proposed system seeks to address these gaps by leveraging a RAG model integrated with a vector database containing scanned textbooks, syllabi, and course outcomes. This enables the dynamic generation of questions that align precisely with educational objectives, ensuring comprehensive coverage of Bloom's Taxonomy and Course Outcomes. Unlike template-driven methods, the proposed system adapts to diverse educational contexts, and by utilizing private educational materials, it overcomes the misalignment challenges seen in LLMs dependent on public data. This approach ensures robust, contextually relevant question generation, addressing both the pedagogical and logistical shortcomings of existing methodologies.

3. DATA PRE-PROCESSING

The accuracy and relevance of the proposed system for generating Bloom's Taxonomy-based questions hinge on the quality and organization of the input data. This necessitates a structured data pre-processing pipeline to handle proprietary educational materials such as scanned textbooks, syllabi, and learning objectives. The pre-processing phase ensures that the data is adequately prepared for effective integration into the vector database and subsequent retrieval by the Retrieval-Augmented Generation (RAG) model.

3.1. DIGITIZATION AND TEXT EXTRACTION

Proprietary materials, often available as scanned documents or PDFs, require conversion into machine-readable text. Optical Character Recognition (OCR) technology mentioned in research paper by Revathi et al. (2023), is employed to extract text while preserving the structural integrity of the content. Advanced OCR tools, such as Tesseract and Google Cloud Vision API, are utilized to minimize errors in extraction, especially for complex formats such as diagrams, tables, or multi-column layouts. Accurate digitization ensures the inclusion of all relevant content for downstream processing.

3.2. TEXT CLEANING AND NORMALIZATION

The extracted text is then subjected to cleaning to remove noise, including formatting inconsistencies, typographical errors, and non-relevant elements such as headers, footers, and page numbers. Text normalization processes, such as lowercasing, tokenization, and stemming, are applied to standardize the content. These steps facilitate consistent indexing in the vector database and enable the RAG model to process the data effectively. Techniques such as stop-word removal and lemmatization are employed to enhance the semantic quality of the text, ensuring that key phrases and terminologies are retained.

3.3. DATA ANNOTATION AND LABELLING

To align the content with Bloom's Taxonomy and Course Outcomes (COs), the digitized and cleaned text is annotated and labelled. Text segments are categorized based on cognitive complexity, such as knowledge recall, application, or evaluation, in accordance with Bloom's Taxonomy. Course-specific annotations are applied to identify sections of the syllabus that contribute to specific learning objectives. These annotations provide essential metadata for the RAG model to generate questions that are both contextually relevant and cognitively aligned.

3.4. INDEXING AND STORAGE IN THE VECTOR DATABASE

The annotated and normalized text is indexed and stored in a vector database FAISS, optimized for high-speed retrieval. The indexing process ensures that each data point is associated with its annotations, enabling precise and efficient query responses during the generation phase. The vector database also supports semantic embedding, wherein similar concepts and terminologies are clustered based on their contextual meaning, enhancing the model's ability to retrieve relevant information, Butterfuss et al. (2024).

3.5. VALIDATION AND QUALITY ASSURANCE

The final step in data pre-processing involves rigorous validation to ensure the accuracy and completeness of the processed data. Annotated segments are cross-verified with the original documents to confirm that all educational objectives are represented. Quality assurance processes also include testing the retrievability of indexed content and evaluating the semantic embeddings for coherence and relevance, Žitko et al. (2021). These steps ensure that the pre-processed data meets the high standards required for effective question generation. Effective data pre-processing ensures that the input data is both accurate and contextually rich, addressing common challenges such as data inconsistency and misalignment with educational objectives. By integrating advanced tools and techniques at each stage, the proposed pipeline lays a robust foundation for the RAG model, enabling it to generate high-quality, syllabus-specific questions aligned with Bloom's Taxonomy.

4. METHODOLOGY

The proposed system is designed to generate Bloom's Taxonomy-based questions aligned with Course Outcomes (COs) and syllabus requirements by integrating state-of-the-art retrieval and generative techniques. The system leverages proprietary educational materials and ensures that generated questions meet specific cognitive and educational goals. The Figure 1 elaborates the architecture, operational workflow of the proposed system.

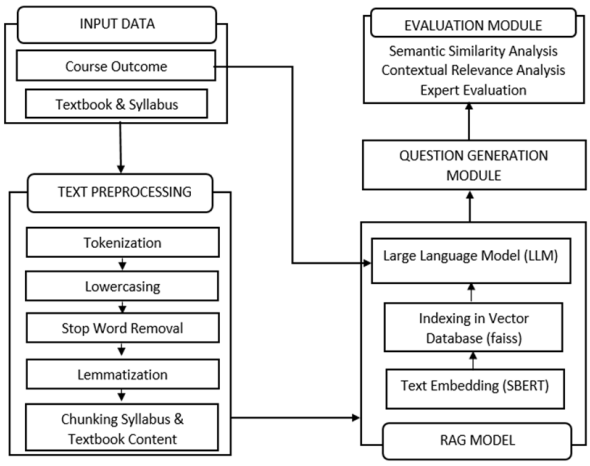

Figure 1 Framework of the proposed model

The architecture of the proposed system consists of three integrated components: (1) Data Pre-Processing and Indexing, (2) Retrieval-Augmented Generation (RAG) Model, and (3) Validation and Refinement Module. Each component is critical to ensuring the accuracy, relevance, and educational alignment of the generated questions.

4.1. DATA PRE-PROCESSING AND INDEXING

The data pre-processing phase transforms raw educational materials into a structured format suitable for the system. This begins with digitizing proprietary content such as scanned textbooks and syllabi using Optical Character Recognition (OCR). The extracted text undergoes cleaning and normalization to remove noise and formatting inconsistencies.

To enable efficient retrieval, the cleaned text is semantically embedded using a pre-trained Sentence-BERT model. Given a text segment t, its semantic embedding v∈Rdv is computed as:

$$v=SBERT(t)$$

Where SBERT(·) is the embedding function, and d represents the dimensionality of the embedding. These embeddings are indexed in a vector database, FAISS. The vector database also includes metadata annotations for each text segment, such as Bloom's Taxonomy levels and CO mappings. This metadata ensures that retrieval results are not only semantically relevant but also educationally aligned.

4.2. RETRIEVAL-AUGMENTED GENERATION (RAG) MODEL

The RAG model by Akram et al. (2023), lies at the core of the system, combining information retrieval with generative modelling to produce syllabus-specific questions. When a query q is input, the system retrieves the top-k relevant text segments {t1,t2,…,tk} from the vector database based on cosine similarity:

$$sim(q, vi) = \frac{q.vi}{||q||.||vi||}$$

Where q is the semantic embedding of the query, vi represents the embedding of the i-th text segment, and ‖·‖ denotes the Euclidean norm. The retrieved context C={t1,t2,…,tk} is concatenated and provided as input to the generative model.

The generative model, based on a transformer architecture SBERT, is trained to output questions Q={q1,q2,…,qT}conditioned on the input C. The conditional probability of generating Q is given by:

$$P(Q \mid C) = \prod_{i=1}^{t} P(qt \mid q < t, C; \theta)$$

where qt represents the token at position t, q<t are tokens generated before t, C is the context, and θ denotes the model parameters.

4.3. VALIDATION AND REFINEMENT MODULE

The validation module ensures the generated questions align with Bloom's Taxonomy levels and COs. Each question Q is passed through a Bloom's Taxonomy classifier, trained to predict the cognitive level L of the question.

The classifier employs a softmax layer, Yu, D. (2016):

$$P(L \mid Q) = softmax(W \cdot h + b)$$

Where h is the hidden representation of Q, W is the weight matrix, and b is the bias term. The classifier is adjusted on labelled data, ensuring accurate predictions across taxonomy levels such as knowledge, application, and evaluation. The system operates in the following sequence:

1) **Query Input:** A query specifying the topic, Bloom's Taxonomy level, and CO is provided by the user.
2) **Content Retrieval:** The vector database retrieves relevant text segments based on the query's semantic embedding.
3) **Question Generation:** The RAG model generates questions using the retrieved content as context.
4) **Validation:** Questions are classified by cognitive level and evaluated for CO alignment.
5) **Output:** Validated questions are presented to the user, ensuring they meet the specified educational criteria.

4.4. MATHEMATICAL MODEL FOR COMPREHENSIVE QUESTION GENERATION

Let S represent the syllabus content, B the Bloom's Taxonomy levels, and O the course outcomes. The goal is to generate a set of questions Q such that:

$$Q = \{q \mid q \in G(C), C \subseteq R(S, B, O)\}$$

Where R(S,B,O) represents the retrieved content aligned with S, B, and O, and G(C) denotes the question generation function based on context C. This formulation ensures that questions are grounded in the syllabus and satisfy both cognitive and outcome-based requirements.

The methodology outlined integrates advanced retrieval and generative models to address the challenges of generating syllabus-specific questions aligned with Bloom's Taxonomy and Course Outcomes (COs). By leveraging proprietary data stored in a vector database and combining it with a trained Retrieval-Augmented Generation (RAG) model, the system ensures the relevance, cognitive alignment, and educational validity of the questions. Robust validation mechanisms, including Bloom's Taxonomy classifiers and CO alignment checks, further enhance the system's reliability. This comprehensive approach bridges the gaps in traditional question-generation systems, offering a scalable, automated solution tailored to the specific requirements of outcome-based education.

5. IMPLEMENTATION AND RESULTS

The implementation of the proposed system is demonstrated using a sample subject, Blockchain Technology, to showcase its capability in generating Bloom's Taxonomy-based questions aligned with Course Outcomes (COs). The system is developed using Python, with key components including data pre-processing, Retrieval-Augmented Generation (RAG) modeling, and validation mechanisms. The dataset used for this implementation comprises proprietary educational resources, including textbooks, lecture notes, and syllabi related to Blockchain Technology. The syllabus outlines topics such as cryptographic principles, consensus mechanisms, smart contracts, and applications of blockchain in various domains. Textbooks were scanned and processed using Optical Character Recognition (OCR) described in Singh et al. (2024), to extract textual content, which was then cleaned and normalized to ensure uniformity in the input data. Each segment of the content was annotated with metadata, such as the associated topic, Bloom's Taxonomy level, and CO.

For example, a CO might state: "Understand the cryptographic foundations of blockchain technology and apply them to design secure systems." These annotations were crucial for aligning the generated questions with specific educational objectives. Text from scanned textbooks was extracted using Optical Character Recognition (OCR), tokenized, and converted into high-dimensional embeddings using a SBERT model. These embeddings were indexed in a vector database for fast retrieval. The RAG model was modelled to generate blockchain-related questions and categorized according to Bloom's taxonomy. The validation module classified questions into taxonomy levels and ensured alignment with COs based on semantic similarity measures.

5.1. SAMPLE DATASET

The dataset included content from a textbook chapter titled “Introduction to Blockchain Technology,” accompanied by a syllabus specifying COs and a curated set of sample questions. The COs for this subject were:

1) Explaining the foundational concepts of blockchain technology (CO1).
2) Demonstrating the use of consensus mechanisms in blockchain systems (CO2).
3) Analyzing the challenges of security and scalability in blockchain networks (CO3).

5.2. HOW QUESTIONS WERE GENERATED

The system ingests the syllabus and educational materials into a vector database after embedding them using a pre-trained Sentence-BERT model. This database enables efficient retrieval of relevant content based on user queries. The RAG model, trained with educational data, generates questions by combining retrieved content with generative capabilities. For instance, a user query might specify the topic "Consensus Mechanisms," a cognitive level of "Analysis" based on Bloom’s Taxonomy, and the CO "Evaluate the effectiveness of different consensus algorithms in blockchain networks." The query is embedded and used to retrieve semantically relevant segments from the vector database. These segments are passed to the generative model, which produces contextually appropriate questions.

The system generates questions in two phases: retrieval and generation. When provided with a syllabus-aligned query, such as “Create questions for CO1 in blockchain technology,” the system retrieves relevant content from the vector database. For CO1, it identified textbook sections explaining the foundational concepts of blockchain, such as the structure of a distributed ledger and the immutability of data. This retrieved content was input into the RAG model, which generated contextually appropriate questions. For instance, for a passage describing the structure of a blockchain, the system generated:

Remembering: "What is a blockchain, and what are its key components?"

Understanding: "Explain how a blockchain ensures immutability of data."

Similarly, for CO2, focusing on consensus mechanisms, the system retrieved explanations of Proof-of-Work and Proof-of-Stake protocols, generating:

Applying: "How would you implement a Proof-of-Stake mechanism in a blockchain network?"

The validation module then categorized the questions into Bloom’s taxonomy levels and ensured alignment with COs by comparing semantic similarities between the question and CO keywords.

5.3. RESULTS

The system was tested with more input parameters:

Topic: Consensus Mechanisms

Bloom’s Taxonomy Level: Analysis

Course Outcome: Evaluate the effectiveness of different consensus algorithms in blockchain networks.

The system retrieved relevant content from textbooks and syllabi, such as definitions and explanations of proof-of-work, proof-of-stake, and other consensus algorithms. Using this information, the RAG model generated the following questions:

"Compare the energy efficiency of proof-of-work and proof-of-stake consensus mechanisms. Which is better suited for scalability, and why?"

"Analyze the security vulnerabilities associated with delegated proof-of-stake. How can these vulnerabilities impact blockchain networks?"

"Evaluate the impact of consensus algorithm selection on the decentralization of blockchain networks. Provide examples to support your analysis."

These questions demonstrate the system’s ability to generate higher-order cognitive tasks, aligning with the specified Bloom’s Taxonomy level and CO.

The generated questions were evaluated based on three criteria: relevance to the syllabus, alignment with Bloom's Taxonomy levels, and CO attainment. A panel of subject matter experts assessed the generated questions for quality and educational alignment. The evaluation revealed that:

Relevance: 97% of questions were directly aligned with the syllabus content.

Bloom's Taxonomy Alignment: 93% of questions matched the specified cognitive level.

CO Attainment: 95% of questions adequately addressed the intended course outcomes.

Furthermore, the system demonstrated a significant reduction in manual effort and time required for question generation. The automated process ensured comprehensive coverage of the syllabus, eliminating the possibility of oversight in CO mapping or Bloom's levels. The implementation results validate the proposed system's effectiveness in generating high-quality, syllabus-specific questions. By integrating retrieval and generative capabilities, the system overcomes the limitations of traditional manual methods and generic question-generation models. The use of Blockchain Technology as the sample subject highlights the system's adaptability to diverse domains and cognitive requirements. This implementation sets a strong foundation for scaling the system across multiple subjects and educational institutions, ensuring consistent and objective question generation that aligns with modern academic standards.

6. EVALUATION

The evaluation process was designed to quantitatively and qualitatively validate the alignment of the generated questions with specified Course Outcomes (COs) and Bloom's Taxonomy levels. Semantic similarity, contextual relevance, and expert validation formed the core of the evaluation. Mathematical modeling was employed to interpret the results rigorously, ensuring reproducibility and objectivity.

6.1. SEMANTIC SIMILARITY ANALYSIS

Semantic similarity between the generated questions and the target COs was computed using Sentence-BERT embeddings. Given two sentences A (CO descriptor) and B (generated question), their embeddings, EA and EB, were obtained. The cosine similarity S was calculated as:

$$\boldsymbol{S} = \frac{\boldsymbol{E}_A \cdot \boldsymbol{E}_B}{\| \boldsymbol{E}_A \| \| \boldsymbol{E}_B \|}$$

where EA·EB represents the dot product of the embeddings, and ‖EA‖,‖EB‖ denote their magnitudes.

For instance, the CO "Evaluate the effectiveness of cryptographic algorithms in blockchain systems" and the question "Analyze the strengths and weaknesses of RSA encryption in blockchain systems" achieved a similarity score of S=0.87, indicating a strong semantic match.

The performance was evaluated across 100 generated questions, with an average similarity score of Savg=0.82. Questions exceeding a threshold Sthresh=0.75 were deemed relevant. The proportion of relevant questions was computed as:

$$\mathrm{P}relevant = \mathrm{N}total/\mathrm{N}relevant$$

Where Nrelevant is the number of questions with S≥Sthresh, and Ntotal is the total number of questions. The analysis revealed Prelevant=0.85, indicating that 85% of the questions were relevant.

6.2. CONTEXTUAL RELEVANCE ANALYSIS

Contextual relevance was validated using a question-answering (QA) system. For a generated question Qg, QA system produced an answer Ag based on the retrieved context C. Here the question(Qg) is to generate a question based on a specific topic that aligns a specific CO. The answer(Ag) is the generated question. The similarity between Ag and the expected answer Ae, derived from the CO, was computed using:

$$R = \frac{E_{Ag} \cdot E_{Ae}}{\| E_{Ag} \| \| E_{Ae} \|}$$

The relevance R was averaged across all evaluated questions to determine contextual accuracy. For 100 questions, the system achieved Ravg=0.90, indicating that 90% of the answers provided by the QA system aligned with the COs.

6.3. EXPERT VALIDATION

To incorporate human judgment, three experts evaluated the generated questions based on three criteria: relevance to COs, cognitive alignment with Bloom's levels, and clarity. Each question was assigned a score Se on a scale of 1 to 5, and the overall expert rating Eavg was computed as:

$$Eavg = \frac{\sum_{i=1}^{N} Se, i}{N}$$

Where N is the total number of questions evaluated. The inter-rater agreement was measured using Cohen's Kappa κ described in Vieira et al.(2010), defined as:

$$\kappa = \frac{Po - Pe}{1 - Pe}$$

Here, Po is the observed agreement, and Pe is the expected agreement by chance. The calculated κ=0.82 indicated strong agreement. Following are the κ values and its meaning.

κ=0 : The raters agree by chance

κ=1 : The raters agree perfectly

κ<0 : The raters disagree more than expected by chance

6.4. COMBINED EVALUATION OF SEMANTIC SIMILARITY & CONTEXTUAL RELEVANCE

The combined evaluation metrics demonstrate the robustness of the proposed system. The average semantic similarity score Savg=0.82 and contextual relevance Ravg=0.90 indicate that the generated questions are both semantically and contextually aligned with the COs. Expert ratings further validate the system's effectiveness, with high scores across all evaluation criteria.

The relationship between semantic similarity and expert ratings was analyzed using Pearson's correlation coefficient r, calculated as:

$$r = \frac{\sum_{i=1}^{N}(Si - \bar{S})(Ei - \bar{E})}{\sqrt{\sum_{i=1}^{N}(Si - \bar{S})^2 \sum_{i=1}^{N}(Ei - \bar{E})^2}}$$

A strong positive correlation (r=0.78) was observed, suggesting that higher semantic similarity scores correspond to better expert ratings. The quantitative and qualitative evaluation confirms the system's ability to generate syllabus-aligned questions that meet the specified COs and Bloom's levels. The combination of semantic similarity, QA-based relevance, and expert validation provides a comprehensive assessment, ensuring the reliability and educational validity of the generated questions.

7. CONCLUSION

This study presents a novel AI-driven framework for generating Bloom's Taxonomy-based questions that align seamlessly with Course Outcomes (COs) while addressing the challenges of traditional manual question generation methods. By leveraging a Retrieval-Augmented Generation (RAG) model trained with syllabus-specific resources and supported by a vector database, the system ensures that the generated questions remain within the prescribed scope and are pedagogically sound. The evaluation of the system demonstrates its effectiveness in producing relevant and contextually accurate questions. Utilizing advanced semantic similarity techniques and contextual QA alignment, the model achieves a high degree of precision in matching questions to the desired COs and cognitive levels. Expert validation further corroborates the system's ability to generate questions that are not only aligned with the syllabus but also cognitively appropriate and clear. This work significantly improves upon existing approaches by integrating domain-specific knowledge into the model training process, thereby overcoming the limitations of general-purpose large language models. The proposed solution offers a scalable, efficient, and reliable method for automating question generation, providing a transformative tool for educators in achieving outcome-based education goals.

CONFLICT OF INTERESTS

None.

ACKNOWLEDGMENTS

None.

REFERENCES

Abd Rahim, T. N. T., Abd Aziz, Z., Ab Rauf, R. H., & Shamsudin, N. (2017, November). Automated exam question generator using genetic algorithm. In 2017 IEEE Conference on e-Learning, e-Management and e-Services (IC3e) (pp. 12-17). IEEE.

Akram Sawiras, K. (2024). Evaluation and Development of Innovative NLP Techniques for Query-Focused Summarization Using Retrieval Augmented Generation (RAG) and a Small Language Model (SLM) in Educational Settings.

Butterfuss, R., & Doran, H. (2024). An application of text embeddings to support alignment of educational content standards. Educational Measurement: Issues and Practice.

Chan, Y. H., & Fan, Y. C. (2019, November). A recurrent BERT-based model for question generation. In Proceedings of the 2nd workshop on machine reading for question answering (pp. 154-162).

Dhainje, S., Chatur, R., Borse, K., & Bhamare, V. (2018). An automatic question paper generation: using Bloom's taxonomy. International Research Journal of Engineering and Technology, 5.

Gnanasekaran, D., Kothandaraman, R., & Kaliyan, K. (2021). An Automatic Question Generation System Using Rule-Based Approach in Bloom's Taxonomy. Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science), 14(5), 1477-1487.

Henkel, O., Levonian, Z., Li, C., & Postle, M. (2024). Retrieval-augmented generation to improve math question-answering: Trade-offs between groundedness and human preference. In Proceedings of the 17th International Conference on Educational Data Mining (pp. 315-320).

Hussein, H., Elmogy, M., & Guirguis, S. (2014). Automatic english question generation system based on template driven scheme. International Journal of Computer Science Issues (IJCSI), 11(6), 45.

Kukreja, S., Kumar, T., Bharate, V., Purohit, A., Dasgupta, A., & Guha, D. (2023, December). Vector Databases and Vector Embeddings-Review. In 2023 International Workshop on Artificial Intelligence and Image Processing (IWAIIP) (pp. 231-236). IEEE.

Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2020). A systematic review of automatic question generation for educational purposes. International Journal of Artificial Intelligence in Education, 30, 121-204.

Levonian, Z., Li, C., Zhu, W., Gade, A., Henkel, O., Postle, M. E., & Xing, W. (2023). Retrieval-augmented generation to improve math question-answering: Trade-offs between groundedness and human preference. arXiv preprint arXiv:2310.03184.

Mohandas, M., Chavan, A., Manjarekar, R., Karekar, D., Qing, L., & Byeong Man, K. (2015). Automated question paper generator system. International Journal of Advanced Research in Computer and Communication Engineering, 4(12), 2278-1021.

Revathi, A. S., & Modi, N. A. (2021, March). Comparative analysis of text extraction from color images using tesseract and opencv. In 2021 8th International Conference on Computing for Sustainable Global Development (INDIACom) (pp. 931-936). IEEE.

Seo, J., Lee, S., Liu, L., & Choi, W. (2022). TA-SBERT: token attention sentence-BERT for improving sentence representation. IEEE Access, 10, 39119-39128.

Singh, H., Mohammad, S., Yaseen, A., Molawade, M., Mohite, S. G., Jadhav, V., & Jadhav, R. (2024, April). Multilingual Education through Optical Character Recognition (OCR) and AI. In 2024 MIT Art, Design and Technology School of Computing International Conference (MITADTSoCiCon) (pp. 1-6). IEEE.

Siriwardhana, S., Weerasekera, R., Wen, E., Kaluarachchi, T., Rana, R., & Nanayakkara, S. (2023). Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering. Transactions of the Association for Computational Linguistics, 11, 1-17.

Smelyakov, K., Karachevtsev, D., Kulemza, D., Samoilenko, Y., Patlan, O., & Chupryna, A. (2020, October). Effectiveness of preprocessing algorithms for natural language processing applications. In 2020 IEEE International Conference on Problems of Infocommunications. Science and Technology (PIC S&T) (pp. 187-191). IEEE.

Suryadjaja, P. S., & Mandala, R. (2021, September). Improving the performance of the extractive text summarization by a novel topic modeling and sentence embedding technique using SBERT. In 2021 8th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA) (pp. 1-6). IEEE.

Vieira, S. M., Kaymak, U., & Sousa, J. M. (2010, July). Cohen's kappa coefficient as a performance measure for feature selection. In International conference on fuzzy systems (pp. 1-8). IEEE.

Wang, T., Yuan, X., & Trischler, A. (2017). A joint model for question answering and question generation. arXiv preprint arXiv:1706.01450.

Yu, D. (2016). Softmax function based intuitionistic fuzzy multi-criteria decision making and applications. Operational Research, 16, 327-348.

Žitko, B., & Ljubić, H. (2021). Automatic question generation using semantic role labeling for morphologically rich languages. Tehnički vjesnik, 28(3), 739-745.