MJoSHT
Malaysian Journal of Science, Health & Technology
[mjosht.usim.edu.my]

USIM

Article

Advanced NLP Techniques for Generating Contextual and Grammatical Arabic Exam Questions

A H Azni[1,2], Farida Ridzuan[1,2], Najwa Hayaati Mohd Alwi [1,2], Sakinah Ali Pitchay[1,2], Zainur Rijal Abd Razak[3], Hanif Ridzwan Ahmad Rodzi[1] and Ahmed A AlSabhany[4]

[1]*Faculty of Science and Technology, Universiti Sains Islam Malaysia, Nilai 71800, Negeri Sembilan, Malaysia.*

[2]*CyberSecurity and Systems Research Unit, Faculty of Science and Technology, Universiti Sains Islam Malaysia, Nilai 71800, Negeri Sembilan, Malaysia.*

[3]*Faculty of Major Language Studies, Universiti Sains Islam Malaysia, Bandar Baru Nilai, 71800 Nilai, Negeri Sembilan, Malaysia.*

[4]*Department of Electronics and Telecommunication Engineering, Daffodil International University Dhaka, Bangladesh.*

Correspondence should be addressed to:

A H Azni; ahazni@usim.edu.my

Article Info
Article history:
Received:19 August 2025
Accepted: 15 January 2026
Published:15 Mac 2026

Academic Editor:
Fauziah Abdul Wahid

Malaysian Journal of Science, Health & Technology

MJoSHT2025, Volume 11, Special Issue on the 5th International Conference on Recent Advancements in Science and Technology (ICoRAST 2025): Responsible Artificial Intelligence – Advancing Science and Technology for Humanity
eISSN: 2601-0003

https://doi.org/10.33102/mjosht.525

Abstract*— This paper outlines the development of an Arabic exam question generator that utilizes advanced Natural Language Processing (NLP) techniques and a comprehensive Arabic corpus. The primary aim is to aid educators in automating the process of crafting exam questions tailored specifically for A1-level Arabic learners. By harnessing the capabilities of NLP, the system integrates sequence-to-sequence (seq2seq) models and template-based methods to generate educationally appropriate questions. The seq2seq models are designed to predict the next word in a sequence, ensuring that the generated questions are natural and contextually fitting. This approach enables the system to produce logically coherent questions that align with the given context. Moreover, the template-based method guarantees grammatical accuracy, which is essential for educational purposes. The templates use as structured guidelines that steer the seq2seq models, ensuring that the questions adhere to proper grammatical rules and structures. A vital aspect of the system is the incorporation of the AraBERT pre-trained model. AraBERT, a transformer-based model customized for Arabic, undergoes fine-tuning with a specifically annotated dataset to adapt it to the task of generating questions from simple Arabic sentences, thereby enhancing its ability to handle the intricacies of the Arabic language. By combining seq2seq models for contextual relevance and template-based methods for grammatical precision, this dual approach effectively addresses the unique challenges associated with Arabic NLP. The richness of Arabic morphology and its syntactic complexity pose significant hurdles for NLP applications. Through the integration of these methodologies, the system ensures that the generated questions are not only contextually relevant but also grammatically correct, making it a valuable tool for educators. In conclusion, the paper discusses an innovative application of advanced NLP techniques and Arabic corpus utilization, providing a robust solution for automated Arabic exam question generation. This system holds significant potential for enhancing the efficiency and effectiveness of language instruction for Arabic learners.

***Keywords*— NLP; Exam Question Generation; Arabic Corpus**

I. INTRODUCTION

The integration of Artificial Intelligence (AI) into the education sector has gained significant momentum in recent years, revolutionizing how instructional content is delivered, personalized, and assessed. Despite this progress, teaching and assessing the Arabic language continues to present notable challenges due to its complex morphology, syntax, and contextual variability. Unlike English and other languages with relatively straight forward grammatical structures, Arabic lacks sufficient AI-powered tools capable of supporting educators in generating diverse, contextually appropriate, and grammatically accurate examination questions. This technological shortfall underscores the pressing need for intelligent systems tailored specifically to Arabic education. As AI continues to expand its role in learning environments, numerous applications often referred to as "intelligent learning tools" [1] have emerged to assist educators in analysing content and designing personalized assessments. These tools aim not only to automate the creation of exam materials but also to align questions with students' cognitive levels, thereby fostering deeper understanding and the development of higher-order thinking skills.

The widespread adoption of AI-driven platforms such as Grammarly, Turnitin, and Duolingo illustrates the growing trust in generative AI across various educational contexts. Tools like Quillionz, for example, utilize machine learning and natural language processing techniques to automatically generate various question types from textual input [2]. However, such tools are predominantly optimized for English, with limited or no support for the Arabic language. As a result, Arabic educators remain largely dependent on manual processes for exam question creation, highlighting a critical gap in AI applications for Arabic language instruction. Addressing this gap requires the development of advanced tools capable of understanding and generating Arabic language structures. The rich linguistic features of Arabic including root-based morphology, diacritics, and contextual variations demand specialized approaches in Natural Language Processing (NLP). The potential of AI to analyse Arabic educational content and generate relevant, context-sensitive questions is substantial and yet underutilized [3].

This paper presents the design and implementation of an AI-powered Arabic exam question generator, which leverages advanced NLP techniques and a curated Arabic corpus. The proposed system aims to streamline the question-generation process, enabling educators to produce diverse and pedagogically sound questions efficiently. Developed as a web-based application, the system incorporates AI-driven linguistic analysis and contextual understanding to ensure that the output aligns with curricular standards while promoting critical thinking among learners. By addressing the limitations of existing tools and focusing on the unique needs of Arabic language education, this research contributes a novel solution to enhance the teaching and assessment experience.

II. RELATED WORKS

Automated exam question generation has significantly transformed educational assessment by enabling the rapid creation of effective, contextually relevant, and pedagogically sound questions. These AI-powered tools are capable of generating various question types including multiple-choice, true/false, and short-answer formats. Thus, streamlining the assessment process for educators. Their adoption is particularly beneficial in higher education and online learning environments, where they reduce the manual workload involved in question creation while ensuring alignment with instructional objectives and the intended cognitive skill levels of learners [4]. This advancement contributes to both the efficiency and effectiveness of assessment practices.

One of the prominent tools in this domain is Quillionz, a next-generation AI-driven platform designed to automate quiz and assessment development. Quillionz leverages a combination of machine learning algorithms and natural language processing (NLP) techniques to generate diverse and high-quality questions from instructional text. While the tool currently demonstrates strong proficiency in English, it also offers preliminary support for Arabic. According to Poonam Jaypuria, Vice President of the Harbinger Group, Quillionz was developed to streamline the question-creation process, enabling educators to generate meaningful and varied assessments quickly [5]. Its intelligent engine employs syntactic, semantic, and template-based methods to transform content into relevant questions, ensuring they are both grammatically sound and pedagogically appropriate [2]. Recent advancements in deep learning have further enhanced the capabilities of question generation systems. Transformer-based models, particularly Bidirectional Encoder Representations from Transformers (BERT) and Generative Pre-trained Transformer (GPT), have emerged as state-of-the-art techniques in this area. BERT processes text bidirectionally capturing contextual relationships from both left-to-right and right-to-left which enables a deeper understanding of word meanings within their surrounding context [6]. This makes BERT especially well-suited for tasks requiring nuanced comprehension, such as question generation. On the other hand, GPT models utilize unsupervised learning to generate human-like text based on input prompts [7]. GPT's strength lies in its generative capabilities, allowing it to create coherent, context-sensitive questions that mirror natural language, making it an effective tool for automated assessment systems.

Other notable platforms include Learnosity and Easygenerator. Learnosity offers a comprehensive suite of tools for developing standards-aligned assessments with AI-enhanced features, including deep analytics, extensive question banks, and seamless integration with various educational technologies [8]. However, new users may experience a learning curve due to the platform's complex interface. In contrast, Easygenerator emphasizes ease of use through its intuitive interface and real-time analytics. It supports multiple question types and integrates efficiently with Learning Management Systems (LMS), making it a practical choice for institutions seeking fast and user-friendly assessment tools [9]. In addition to transformer models, Long Short-Term Memory (LSTM) networks a class of recurrent neural networks (RNNs) have also been widely adopted in question generation. LSTM networks are capable of capturing long-term dependencies in sequential data, which is crucial for

understanding extended contexts in text passages [10]. This enables the generation of coherent and contextually relevant questions, particularly when the necessary information is distributed across several sentences or paragraphs.

Finally, the Transformer architecture, introduced by Vaswani et al. [11], has become a foundational model in modern NLP. Utilizing a self-attention mechanism, transformers weigh the relevance of each word in a sentence relative to others, allowing for a deep and holistic understanding of context. This mechanism has proven highly effective in generating high-quality questions that reflect complex linguistic relationships [12][13]. Their scalability and adaptability also support the development of customized question-generation systems across various domains and languages. Despite these advancements, there remains a considerable gap in tools specifically tailored for Arabic exam question generation. Most existing systems are optimized for English or other widely spoken languages, lacking the morphological and syntactic adaptability required to handle the intricacies of Arabic. This limitation underscores the need for Arabic-focused AI solutions particularly those employing advanced NLP and deep learning models to support educators in generating high-quality, contextually appropriate assessment content for Arabic learners.

III. METHODOLOGY

This section outlines the design diagrams of the Arabic exam question generator system, as shown in Figure 1. The system leverages advanced Natural Language Processing (NLP) techniques and an Arabic corpus to automate the creation of Arabic exam questions, thereby simplifying the question development process. It also describes the system's core functionality, including the development of a web-based platform and its integrated NLP model.

Figure 1. Design Flow for Arabic Question Generator

A. NLP Model Training Design

The primary objective of this paper is to develop an NLP model capable of automating the generation of Arabic exam questions. Before lecturers can utilize the system for question creation, the model must operate effectively to fulfill user expectations. Figure 2 illustrates the development workflow of the NLP model designed for this purpose.

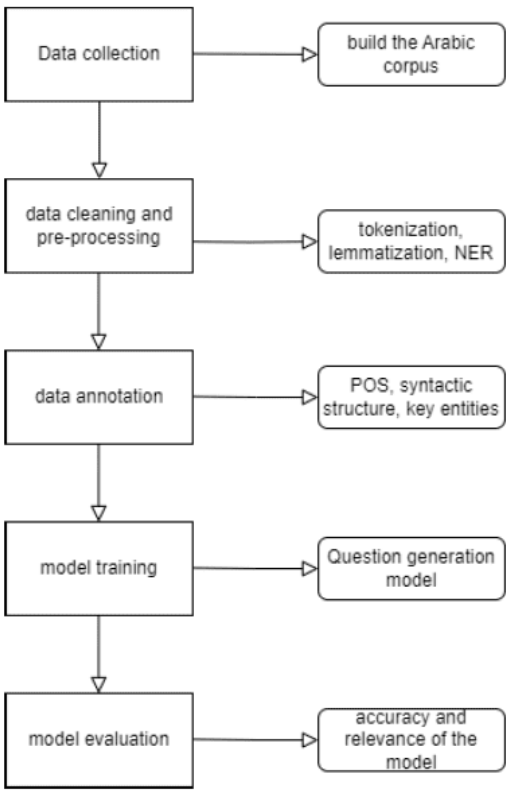

Figure 2. Flow diagram for NLP model development

The process begins with the collection of relevant materials to build an Arabic exam question corpus. This is followed by data cleaning and preprocessing, which aim to remove irrelevant content and retain only the information necessary for the model, specifically targeting the A1 proficiency level. Key preprocessing tasks include tokenization, lemmatization, stemming, and Named Entity Recognition (NER) [14]. After refining the dataset, the annotation phase begins, where linguistic features are labelled to ensure that the vocabulary and sentence structures are appropriate for A1-level learners. This involves part-of-speech tagging, syntactic analysis, and identification of key entities.

Once the data has been prepared, the next stage is model training. This step focuses on developing a question-generation model tailored to the needs of A1-level Arabic students, using a combination of Sequence-to-Sequence (Seq2Seq) and template-based approaches. Following model training, a thorough testing and evaluation phase is conducted. Feedback from lecturers and students plays a vital role in assessing the model's performance particularly in terms of vocabulary usage, grammatical accuracy, and difficulty level.

This feedback informs an iterative process of retraining and fine-tuning, ensuring the model not only meets educational standards but also integrates seamlessly with the web application. Ultimately, this process aims to deliver a reliable system capable of generating high-quality, relevant exam questions for beginner-level Arabic learners.

B. Data Collection and Pre-Processing

To support the generation of Arabic exam questions, a comprehensive corpus specifically tailored to A1-level Arabic proficiency will be compiled. This process involves sourcing materials from beginner-level textbooks and online resources designed for novice learners. An automated web-scraping tool, Scrapy, will be utilized to extract data from educational websites that offer Arabic learning content, with a particular focus on sample questions. Once collected, the corpus will undergo a thorough cleaning and preprocessing stage to ensure alignment with A1-level requirements. This includes the removal of complex vocabulary and advanced grammatical structures that exceed beginner proficiency. Figure 3 illustrates an example of Arabic text preprocessing.

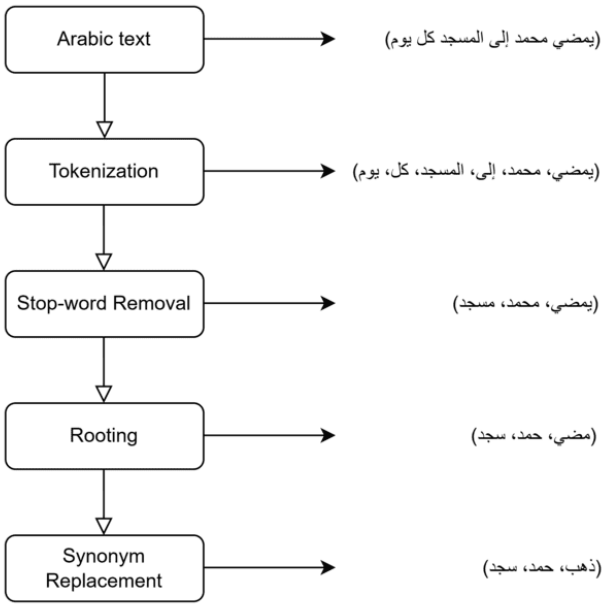

Figure 3. Pre-processing steps of an Arabic text [15]

In the example, the flowchart uses synonym replacement to replace the word with the same meaning to standardize vocabulary. For example, the word "مضي " (went) is replaced with its synonym "ذهب " (go) in the final step. This process ensures that different words that represent the same meaning will be treated equally. Figure 4 depicts the proposed flow chart for Arabic text pre-processing at the A1 proficiency level.

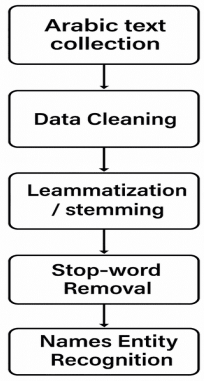

Figure 4. Flow Chart of Data Cleaning and Pre-processing for A1 Proficiency Level

To enhance comprehension, Table I detailed out the workflow demonstrated in Figure 3 and Figure 4 above.

TABLE I. DATA CLEARNING AND PRE-PROCESSING

Arabic text	**يذهب محمد إلى المدرسة كل يوم**
Tokenization	["يذهب"، "محمد"، "إلى"، "المدرسة"، "كل"، "يوم"]
Lemmatization or Stemming	["ذهب"، "محمد"، "إلى"، "درس"، "كل"، "يوم"]
Stop-word removal	["يذهب"، "محمد"، "المدرسة"]
Named Entity Recognition	Identify "محمد " as a person and "المدرسة " as a place.
Question generation	" من يذهب إلى المدرسة؟ " (Who goes to the school?)

C. NLP Development Model

The objective of this section is to build an initial NLP model for creating introductory-level questions in Arabic. The model will utilize transformer architectures like BERT and will undergo fine-tuning using a specifically annotated dataset. Refer to Figure 5 for the flowchart outlining this section.

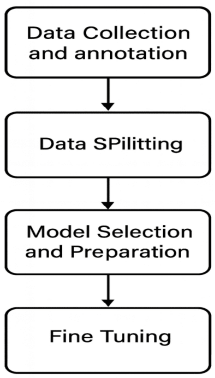

Figure 5. Flow chart of Initial NLP Model Development

D. Data Collection and Annotation

A corpus of simple Arabic sentences that are suitable for A1-level students will be collected and prepared. The source is from learning language resources or simple news articles. To make sure the sentences are appropriate for A1 level students, basic vocabulary and simple sentence structure are focused. Furthermore, the data collection will undergo a pre-processed phase to ensure the collected data are cleaned from unrelated noise.

The sentence will be annotated to highlight key linguistic features. The proposed annotations are Part of speech (POS) tagging, syntactic structure and key entities for question generation. Below is the example for every proposed

annotation in Table II, Table III and Table IV. The sentence used to show an example for annotation: يذهب محمد إلى المدرسة

TABLE III. POS TAGGING

Word	Translation	POS
يذهب	going	verb
محمد	Muhammad	Noun-proper name
إلى	to	preposition
المدرسة	school	noun

TABLE IIIII. SYNTATIC STRUCTURE

Verb Phrase (VP): Verb (V): Noun Phrase (NP):	يذهب محمد يذهب محمد
Prepositional Phrase (PP): Preposition (P): Noun Phrase (NP):	إلى المدرسة إلى المدرسة

TABLE IV. KEY ENTITIES

Key Entities	
Person	محمد
Activity	يذهب
Destination	المدرس

Possible questions can be derived from the annotation above as follows:

- Who goes to the school? - من يذهب إلى المدرسة؟
- Where does Muhammad go? - إلى أين يذهب محمد؟

Full annotated sentence example is shown in Table V.

TABLE V. ANNOTED SENTENCES

POS	"يذهب /V محمد /N إلى /P المدرسة /N"
Syntactic structure	"[VP يذهب [NP محمد]] [PP إلى [NP المدرسة]]"
Key entities	"يذهب /Activity محمد /Person إلى /Preposition المدرسة /Destination"

This comprehensive annotation provides a detailed understanding of the sentence structure, grammatical roles, and key elements, virtualizing the training of an NLP model to generate suitable questions for A1-level Arabic students.

1) Data Splitting

The dataset for this study is meticulously divided into three distinct subsets to ensure optimal model training, validation, and evaluation. Specifically, 70% of the data is allocated to the training set, which serves as the primary source for the model to learn and identify patterns. This substantial portion of the data enables the model to develop a comprehensive understanding of various sentence structures and contexts inherent in Arabic language questions. By exposing the model to a wide array of examples, the training set plays a crucial role in establishing the foundational capabilities of the NLP system [16].

The remaining data is split equally between a validation set and a test set, each comprising 15% of the overall dataset. The validation set is utilized during the training phase to fine-tune hyperparameters and validate the model's performance, ensuring that the system is not overfitting and can generalize well to unseen data. The test set, on the other hand, is reserved for the final evaluation of the model post-training. By assessing the model on this independent subset, we can objectively measure its accuracy and effectiveness in generating Arabic exam questions. This careful partitioning of the dataset helps maintain a balanced representation of sentence structures and contexts across all subsets, thereby enhancing the robustness and reliability of the model's performance [17].

2) Model Selection and Preparation

The focus of this section is the method on how to develop and generate questions. The proposed method uses sequence-to-sequence (seq2seq) models and template-based generation techniques to create questions suitable for A1-level Arabic students.

a. Sequence-to-sequence (seq2seq) models

Seq2Seq models are designed to generate questions by predicting the next word in a sequence, given the context of the previous words [18]. This approach is useful for generating natural and contextually relevant questions. The model will be trained on the Arabic corpus to learn the structure and vocabulary appropriate for A1-level Arabic students. The encoder processes the input sequence like a sentence and converts it into a fixed-length context vector, which is an abstract representation of the input sequence. For example, the the input sentence "أذهب إلى المدرسة كل يوم " (I go to school every day), the encoder generates a context vector representing this information.

The decoder takes the context vector from the encoder and generates the output sequence which is a question one word at a time. It predicts the next word in the sequence based on the previous words and the context vector. For example, using the context vector from the input sentence, the decoder might generate the question "أين تذهب كل يوم؟ " (Where do you go every day?). Seq2Seq models are trained on pairs of input and output sequences. In the context of question generation, the input could be a statement, and the output could be a corresponding question. During training, the model learns to map the structure and meaning of the input sequence to the appropriate output sequence.

b. Template-Based Generation Techniques

Template-based generation relies on predefined templates to create structured questions. This method ensures that the questions are grammatically correct and suitable for beginners, especially for A1 level students [19]. A set of templates for common question structures are shown below.

1. "ماذا تفعل في [time]؟" (What do you do in [time]?)
2. "أين تذهب؟ " (Where do you go [time]?)
3. "من؟ " (Who [verb]?)
4. "هل هذا البيان صحيح أو خطأ؟ " (Is this statement true or false: [statement]?)

These templates have placeholders that can be filled with identified entities and relationships.

3) Fine-tuning

For fine-tuning, the pre-trained Bidirectional Encoder Representations from Transformers (BERT) model is selected. The BERT-base-Arabic model, or AraBERT, will be included for further fine-tuning. The Bert tokenizer will be used to convert text into token IDs, ensuring that the tokens are padded or truncated to a uniform length for batch processing.

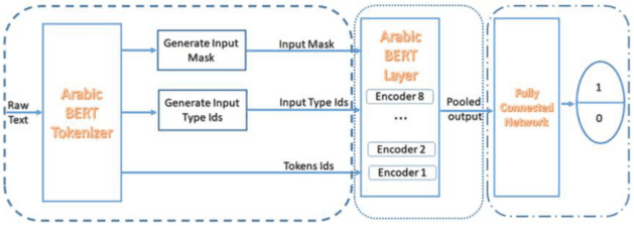

Figure 6. Arabic BERT model architecture [20]

Figure 6 shows the AraBERT model architecture. BERT is a transformer-based model that is designed to understand the context of words in a sentence by looking at both words before and after the targeted word. BERT is initially pre-trained on a large corpus of text, learning to predict masked words and the next sentence in a sequence. BERT is then fine-tuned on a specific task, such as generating questions for A1 learners in Arabic, using a smaller, annotated dataset. BERT models like AraBERT have been specifically pre-trained on large Arabic corpora, capturing the nuances of the Arabic language. Fine-tuning AraBERT involves using the specific annotated dataset to adapt the model to understand and, for example, generating questions based on simple Arabic sentences. The training data primarily consists of learners at the A1 proficiency level, which could limit the diversity of language structures, topics, and difficulty levels represented. This lack of variety could potentially result in biases, such as a skew toward specific language patterns or subject areas, affecting the generalizability of the generated questions.

IV. QUESTION GENERATION PROCESS DESIGN

The following section outlines the methodology for creating and formulating questions. The method involves integrating sequence-to-sequence (seq2seq) models and a template-based generation technique to produce questions that are appropriate for A1-level students. This approach offers a comprehensive and organized overview of the operational process of the Arabic Exam Question Generator system.

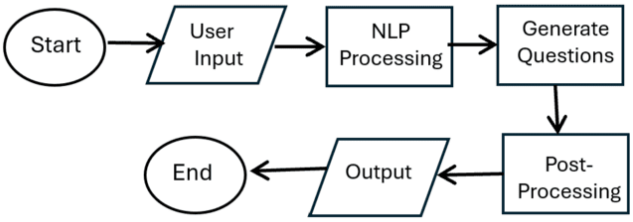

Figure 7. Flowchart of the Arabic Exam Question Generator

Figure 7 shows a proposed flowchart of the system. The process of generating Arabic exam questions begins with user input, where the user enters text in Arabic specifying the type of question they wish to generate. This ensures that the NLP model, designed specifically for Arabic, can accurately process the text and generate relevant questions. Once the input is provided, the system utilizes advanced NLP techniques to analyse the grammatical structure, context, and key elements of the text. A pre-trained Arabic NLP model facilitates this processing, ensuring a thorough understanding of the input [21].

Following the analysis, the system generates questions that align with the processed text, ensuring they are contextually appropriate and suitable for A1-level Arabic students. The generated questions are then classified into different types, such as multiple-choice, fill-in-the-blank, and short answer, to cater to various exam formats. In the post-processing stage, the system reviews and refines the questions, making necessary adjustments to ensure they are grammatically correct and contextually accurate. Finally, the system presents the refined set of questions to the user, displaying them on a webpage for easy access.

A. Arabic Question Generator Web Development

The web-based Arabic Question Generator has two main functions which are the login interface and question generator interface. The login interface in Figure 8 is designed to ensure that only teachers or authorized users can access the system to generate exam questions. To gain access, users must enter their username and password, providing a secure gateway that protects the integrity and confidentiality of the educational content. This security measure is crucial for maintaining control over the question-generation process and ensuring that only qualified individuals can utilize the system's capabilities [22]. By implementing this secure login mechanism, the system prevents unauthorized access and potential misuse of the question generator, safeguarding the educational materials from tampering or leakage. It ensures that only vetted and responsible educators can create and modify exam questions, thereby upholding the quality and integrity of the assessments. Furthermore, this security feature supports compliance with institutional policies and educational standards, reinforcing the system's reliability and trustworthiness. Ultimately, the robust login interface is a vital component in protecting the educational environment and maintaining the credibility of the generated examination content.

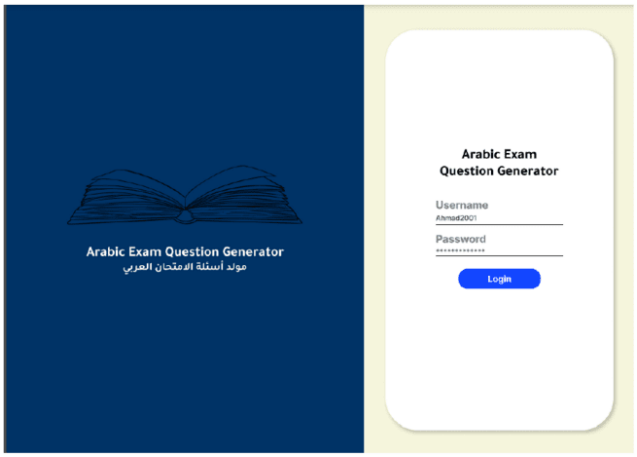

Figure.8. Login interface design

Once logged in, users are directed to a page where they can generate questions. On this page, users can specify the type of question they wish to create and the number of questions required. This flexibility allows educators to tailor the questions to their specific needs and the level of their students [23]. Additionally, users have the option to export the generated questions, create new ones, or delete any existing questions. These functionalities offer a comprehensive and user-friendly experience, enabling educators to efficiently manage and customize their exam content. Figure 9 illustrates the design of the question generator interface, highlighting its straightforward and functional layout.

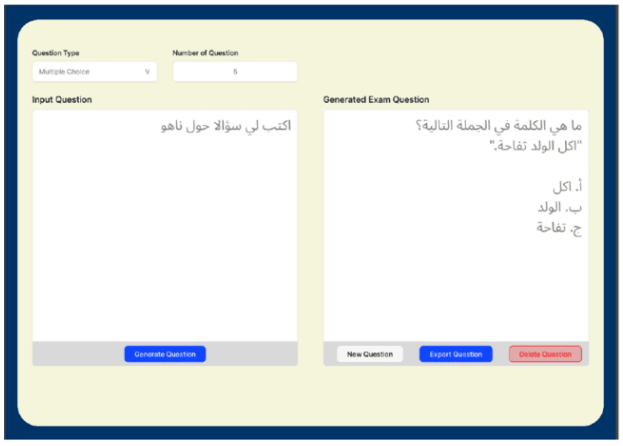

Figure 9: Question generator interface

In this study, the accuracy and effectiveness of the proposed method are not fully addressed due to the limited data available. Future research will focus on evaluating the precision of the questions generated by the system and assessing its overall functionality. Additionally, further testing will aim to refine the algorithm, enhance its adaptability, and ensure its robustness across diverse scenarios. These efforts will contribute to improving the system's performance and offer a more thorough evaluation of its capabilities. However, the findings of this study may not be directly applicable to more advanced learners, as the focus was solely on A1-level proficiency. This limitation suggests that the broader applicability of the system may be restricted, underscoring the need for future investigations that involve learners across a range of proficiency levels to provide a more comprehensive understanding of the system's effectiveness and potential for generalization.

V. TESTING AND EVALUATION

This section presents a comprehensive evaluation of the Arabic Exam Question Generator System, focusing on the User Acceptance Testing (UAT). The evaluation aimed to assess the system's usability, accuracy, responsiveness, and overall effectiveness in fulfilling its primary function, supporting educators in generating high-quality Arabic exam questions. Feedback was gathered via structured user surveys administered to a group of ten (10) users, including lecturers and academic staff. Each question in the survey was measured using a 5-point Likert scale, and the results are detailed and discussed below.

According to v 10, 90% of respondents rated the system as extremely user-friendly, while the remaining 10% rated it as user-friendly. These results clearly indicate that users found the interface intuitive and easy to navigate, even on their first use.

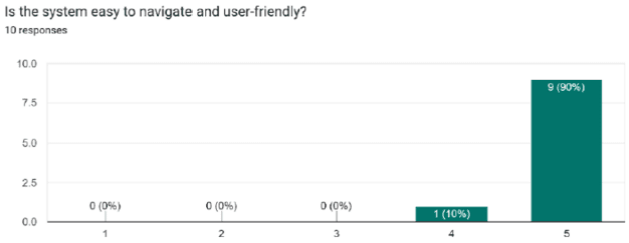

Figure 10. System Navigation and User-Friendliiness Scale

The absence of any neutral or negative responses suggests that the design choices such as menu layout, input fields, and user feedback prompts were well-aligned with user expectations. Such high satisfaction is vital for encouraging repeat usage and reflects the system's accessibility for a wide range of users, including those less experienced with educational technology platforms.

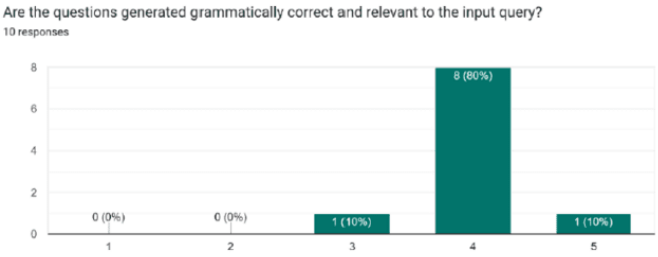

Figure 11. Accuracy and Relevance of Generated Questions Scale

Figure 11 demonstrates that 80% of users rated the generated questions as grammatically correct and contextually appropriate, indicating strong alignment with standard Arabic language rules and curriculum-based content. An additional 10% rated the questions as neutral, and another 10% rated them as highly relevant, further validating the system's core functionality. Although most feedback was positive, the small percentage of neutral responses highlights the importance of ongoing refinement, particularly in semantic understanding

and context-aware phrasing. Improvements in these areas could elevate the system's reliability in handling complex question structures or diverse curriculum requirements.

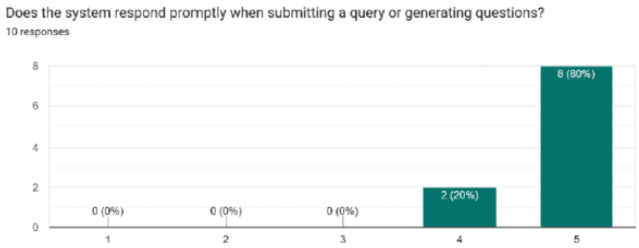

Figure 12. System Responsiveness Scale

Responsiveness is a key determinant of user satisfaction, especially for systems expected to deliver outputs in real-time. As reflected in Figure 12, 80% of users found the system highly responsive, with the remaining 20% rating it as responsive. These Fig.s affirm the robustness of the system's backend performance and its ability to process queries without noticeable delays. From a technical perspective, this suggests that server-side logic and database interactions are optimized, contributing positively to overall usability.

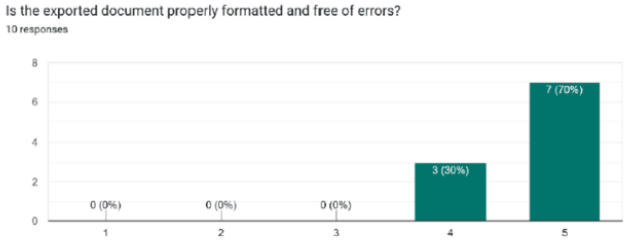

Figure 13. Exported Document Formatting Scale

Figure 13 shows that 70% of respondents agreed that the exported documents were properly formatted and free from typographical or structural issues, while 30% noted minor inconsistencies. This level of satisfaction validates the system's utility for academic use, particularly for printing or uploading exam papers. Nonetheless, the feedback indicates that formatting mechanisms such as font alignment, margin control, and section labelling can be further polished to meet institutional standards more consistently. Future iterations may include a live document preview feature or customizable formatting options for greater flexibility.

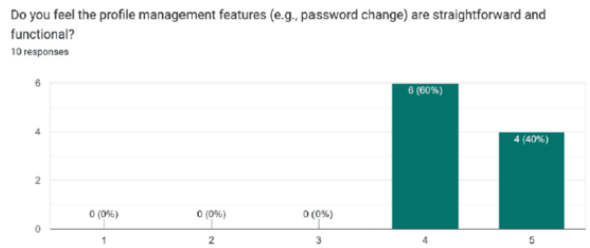

Figure 14. Profile Management Functionality Scale

As indicated in Figure 14, 60% of users found the profile management features, such as password updates and access control, to be straightforward and functional, while 40% rated them as highly functional. These findings suggest that users were able to manage their accounts without technical support, which enhances system autonomy and reduces administrative overhead. However, the system could benefit from adding features like multi-factor authentication, activity logs, or customizable profile settings to align with best practices in security and personalization.

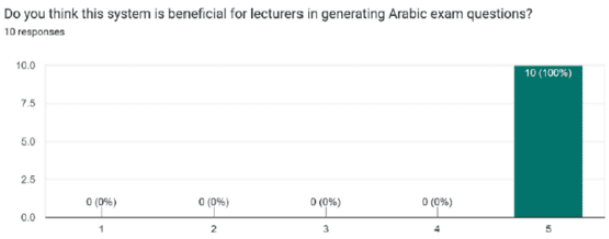

Figure 15. Profile Management Functionality Scale

Figure 15 demonstrates unanimous agreement 100% of users strongly believe the system is highly beneficial for lecturers, particularly in automating the generation of Arabic exam questions. This reflects the system's success in meeting its primary objective: reducing time and cognitive load in exam preparation. Such positive reception also indicates the system's potential for scalability and broader adoption in other academic institutions or for other language subjects.

The results of the UAT indicate that the Arabic Exam Question Generator System performs exceptionally well across all key functional areas. The high satisfaction scores across usability, responsiveness, and relevance affirm the system's readiness for deployment in real-world academic settings. At the same time, minor areas for improvement such as document formatting and advanced profile settings were identified, offering a pathway for iterative refinement. Overall, the evaluation underscores the system's practicality, its alignment with user needs, and its contribution to enhancing educational efficiency through smart automation.

VI. LIMITATION AND FUTURE WORKS

Despite the promising outcomes of this study, several limitations warrant critical consideration. The reliance on a curated A1-level Arabic corpus introduces potential content and domain bias, as the dataset is primarily derived from beginner textbooks and online materials. This lack of linguistic diversity may restrict the system's ability to generate questions that reflect authentic real-world Arabic usage or higher proficiency levels, a concern consistent with earlier corpus-driven NLP research. Another limitation concerns scalability. The system has been optimized specifically for A1 learners, but its performance across more advanced proficiency levels, regional dialects, or domain-specific Arabic (e.g., legal, technical, or religious texts) remains untested. Expanding the system to these contexts requires larger, heterogeneous corpora and robust evaluation mechanisms. Furthermore, the computational cost associated with fine-tuning and deploying large-scale transformer models

such as AraBERT presents challenges for institutions with limited resources, echoing broader scalability issues reported in deep learning applications [6], [11].

Over-reliance on algorithmically generated questions could reduce educator oversight, potentially leading to pedagogical misalignment or reinforcement of unintended biases. Concerns related to data privacy and security also emerge, particularly if future versions incorporate learner-generated data or interaction logs, aligning with ongoing debates on AI in education and responsible data use [21]. To maintain fairness and educational integrity, it is essential that such systems function as decision-support tools rather than substitutes for human educators.

Future work will address these limitations by expanding the corpus to include more diverse, balanced, and multi-dialectal datasets [17], developing bias detection and mitigation pipelines during training [13], and integrating explainability mechanisms to enhance transparency [12]. Incorporating educator-in-the-loop frameworks will help preserve the pedagogical role of teachers, ensuring that the system complements rather than replaces human judgment. Pilot implementations across diverse institutions and learner groups will provide further validation of the system's robustness, adaptability, and scalability, ultimately strengthening its potential for real-world educational deployment.

VI. CONCLUSIONS

In conclusion, this paper presents a pioneering application of advanced NLP techniques combined with a comprehensive Arabic corpus to automate the generation of exam questions tailored for A1-level Arabic learners. By leveraging seq2seq models for contextual coherence and template-based methods for grammatical accuracy, the system adeptly navigates the complexities of Arabic morphology and syntax. The integration of the AraBERT pre-trained model, fine-tuned for this specific task, further enhances the system's capability to generate contextually relevant and grammatically precise questions. This dual approach not only addresses the unique challenges of Arabic NLP but also provides a valuable tool for educators, significantly improving the efficiency and effectiveness of language instruction. The innovative methodologies outlined in this paper underscore the potential for NLP to transform educational practices, offering robust solutions for automated exam question generation in the Arabic language. The implications of this work extend beyond just question generation; it provides a scalable solution that can be implemented across Arabic-speaking regions, enhancing educational accessibility and resource availability. Future work could include pilot programs in educational institutions, integrating the tool into Learning Management Systems (LMS) and refining it for a broader range of proficiency levels. By doing so, this system can not only assist in language learning but also support educators in creating contextually appropriate, grammatically accurate assessments, thereby improving the overall quality of Arabic language education.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this paper.

ACKNOWLEDGEMENT

This research is funded by the Universiti Sains Islam Malaysia (USIM) Research Grant under research grant no: PPPI/USIM/FST/USIM/16924.

REFERENCES

[1] Hwang, Gwo-Jen, Haoran Xie, Benjamin W. Wah, and Dragan Gašević. "Vision, challenges, roles and research issues of Artificial Intelligence in Education." Computers and Education: Artificial Intelligence 1 (2020): 100001. https://doi.org/10.1186/s40537-022-00625-z

[2] Bahy, Mazen. "Comparative analysis of Machinegenerated questions (Quillionz) and Human-generated questions." (2020).

[3] Basha, M. John, S. Vijayakumar, J. Jayashankari, Ahmed Hussein Alawadi, and Pulatova Durdona. "Advancements in natural language processing for text understanding." In E3S Web of Conferences, vol. 399, p. 04031. EDP Sciences, 2023. https://doi.org/10.1051/e3sconf/202339904031

[4] Thotad, Puneeth, Shanta Kallur, and Sukanya Amminabhavi. "Automatic question generator using natural language processing." Journal of Pharmaceutical Negative Results (2022): 2759-2764.

[5] Vakaliuk, Tetiana A., Oleksii V. Chyzhmotria, Svitlana O. Didkivska, and Illia Linevych. "Development of a web service for creating tests based on text analysis using natural language processing technologies." International Journal of Research in E-learning 9, no. 2 (2023): 1-22. https://doi.org/10.31261/IJREL.2023.9.2.04

[6] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[7] Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. "Improving language understanding by generative pre-training." (2018).

[8] Tomberg, Vladimir, Pjotr Savitski, Pavel Djundik, and Vsevolods Berzinsh. "Design and development of IMS QTI compliant lightweight Assessment delivery system." In Technology Enhanced Assessment: 19th International Conference, TEA 2016, Tallinn, Estonia, October 5-6, 2016, Revised Selected Papers 19, pp. 159-170. Springer International Publishing, 2017. https://doi.org/10.1007/978-3-319-57744-9_14

[9] Jones, Heather M. "Using innovative technologies to increase student engagement in an online Anatomy and Physiology course." The FASEB Journal 33, no. S1 (2019): 598-23. https://doi.org/10.1096/fasebj.2019.33.1_supplement.598.23

[10] Sherstinsky, Alex. "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network." Physica D: Nonlinear Phenomena 404 (2020): 132306. https://doi.org/10.1016/j.physd.2019.132306

[11] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." Advances in neural information processing systems 30 (2017). https://doi.org/10.48550/arXiv.1706.03762

[12] Ruiz-Rojas, Lena Ivannova, Patricia Acosta-Vargas, Javier De-Moreta-Llovet, and Mario Gonzalez-Rodriguez. "Empowering education with generative artificial intelligence tools: Approach with an instructional design matrix." Sustainability 15, no. 15 (2023): 11524. https://doi.org/10.3390/su151511524

[13] Maulud, Dastan Hussen, Siddeeq Y. Ameen, Naaman Omar, Shakir Fattah Kak, Zryan Najat Rashid, Hajar Maseeh Yasin, Ibrahim Mahmood Ibrahim, Azar Abid Salih, Nareen OM Salim, and Dindar Mik Ahmed. "Review on natural language processing based on different techniques." Asian Journal of Research in Computer Science 10, no. 1 (2021): 1-17. https://doi.org/10.9734/ajrcos/2021/v10i130231

[14] Gumaste, Priti, Shreya Joshi, Srushtee Khadpekar, and Shubhangi Mali. "Automated Question Generator System Using NLP Libraries." International Research Journal of Engineering and Technology (IRJET) 7, no. 6 (2020): 4568-4572.

[15] Menai, Mohamed El Bachir. "Detection of plagiarism in Arabic documents." International Journal of Information Technology and Computer Science 10, no. 10 (2012): 80-89. https://doi.org/10.5815/ijitcs.2012.10.10

[16] Marie-Sainte, Souad Larabi, Nada Alalyani, Sihaam Alotaibi, Sanaa Ghouzali, and Ibrahim Abunadi. "Arabic natural language processing and machine learning-based systems." IEEE Access 7 (2018): 7011-7020. https://doi.org/10.1109/ACCESS.2018.2890076

[17] Ali, Abbas Raza, Muhammad Ajmal Siddiqui, Rema Algunaibet, and Hasan Raza Ali. "A large and diverse Arabic corpus for language modeling." Procedia Computer Science 225 (2023): 12-21. https://doi.org/10.1016/j.procs.2023.09.086

[18] Keneshloo, Yaser, Tian Shi, Naren Ramakrishnan, and Chandan K. Reddy. "Deep reinforcement learning for sequence-to-sequence models." IEEE transactions on neural networks and learning systems 31, no. 7 (2019): 2469-2489. https://doi.org/10.1109/TNNLS.2019.2929141

[19] He, Xiao, Tian Zhang, Minxue Pan, Zhiyi Ma, and Chang-Jun Hu. "Template-based model generation." Software & Systems Modeling 18 (2019): 2051-2092. https://doi.org/10.1007/s10270-017-0634-5

[20] Chouikhi, Hasna, Hamza Chniter, and Fethi Jarray. "Arabic sentiment analysis using BERT model." In Advances in Computational Collective Intelligence: 13th International Conference, ICCCI 2021, Kallithea, Rhodes, Greece, September 29–October 1, 2021, Proceedings 13, pp. 621-632. Springer International Publishing, 2021. https://doi.org/10.1007/978-3-030-88113-9_50

[21] Alduailej, Alhanouf, and Abdulrahman Alothaim. "AraXLNet: pre-trained language model for sentiment analysis of Arabic." Journal of Big Data 9, no. 1 (2022): 72. https://doi.org/10.1186/s40537-022-00625-z

[22] Al Masri, Ahmad Mustafa Ali, Muhammad Suzuri Hitam, Wan Nural Jawahir Hj Wan Yussof, and Atallah Al-Shatnawi. "Novel Algorithm for Baseline Detection of Offline Arabic Handwritten Text Recognition." Journal of Advanced Research in Applied Sciences and Engineering Technology 37, no. 1 (2024): 56-68. https://doi.org/10.37934/araset.37.1.5668

[23] Ab Halim, A. H., Ridzuan, F., Zakaria, N. H., Zakaria, A. A., Mohd Alwi, N. H., Ali Pitchay, S., & Az-Zuhar, I. (2024). SAKTI©: Secured Chatting Tool Through Forward Secrecy. Journal of Advanced Research in Applied Sciences and Engineering Technology, 49(1), 54–62. https://doi.org/10.37934/araset.49.1.5462