IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 5, May 2026

DOI: 10.17148/IARJSET.2026.13503

Smart Question Paper Generation System Using NLP

Gavini Venkateswari[1], Redrouthu Lavanya[2], Bheemineni Kavya Sudha[3], Mandava Divya [4]

Department of Computer Science and Engineering, Bapatla Women's Engineering College[1]

Department of Computer Science and Engineering (AI & ML), Bapatla Women's Engineering College[2]

Department of Computer Science and Engineering (AI & ML), Bapatla Women's Engineering College[3]

Department of Computer Science and Engineering (AI & ML), Bapatla Women's Engineering College[4]

Abstract: The process of designing the examination papers is quite lengthy, biased, and time-consuming. The current research introduces a novel solution to create automatic examination papers based on the syllabus documents in the PDF format. The system uses PyMuPDF for extracting information and processing unstructured text using state-of-the-art Natural Language Processing tools. The generator makes use of a transformer neural network model named Flan-T5 which can produce multiple- choice questions (MCQs) along with contextually appropriate distractors and descriptive long-answer questions. The system also incorporates a login module to ensure secure access and provides the option of exporting the question papers in TXT and PDF formats. According to experimental results, the system shows a remarkable improvement in the speed of generating questions and saves almost 85 percent of the time as compared to the conventional technique. The experiments also confirm the quality of the system as far as coherence and grammaticality of the generated questions are concerned.

Keywords: Natural Language Processing, Automatic Question Generation, Transformer Models, PDF Text Extraction, Educational Technology

I. INTRODUCTION

A. Background and Context

Advances in digital technologies have revolutionized the world of education, bringing about intelligent systems to automate administrative and instructional activities. The automatic generation of questions papers, however, still constitutes one of the underexploited areas, since assessing student performance is one of the most important stages in the teaching process. Examination papers allow for analyzing whether students have successfully assimilated all theoretical concepts, along with developing their analytical skills. Thus, preparing well- balanced and informative question sets is a laborious activity requiring thorough selection of appropriate content and covering all syllabus topics at the same time.

Traditionally, teachers have been engaged in the tedious and mentally straining task of selecting materials, such as textbook chapters, lectures, and other sources of content, to create question sets. Considering how time-bound educational environments are, creating question sets can take quite a bit of time, which could otherwise be dedicated to teaching activities. Moreover, in today's context, there are numerous digital resources available to educators, such as PDF documents containing relevant content.

B. Problem Definition

Even with the wealth of digital content available for use, question paper generation is still a lengthy procedure that involves manual analysis of resources and selecting suitable material for questions. As such, the first issue concerns the amount of time that should be spent to find relevant questions that would cover all topics and present enough diversity.
Secondly, since manual procedures cannot provide objectivity, there might be a case of some topics being more emphasized than others, thus distorting student perceptions of the subject and undermining fairness.
Thirdly, since all tasks are generated manually, without any guidelines, it becomes quite hard to control the level of difficulty of tasks. Therefore, achieving balance among all sets of questions proves to be rather problematic.
Finally, using unstructured documents, such as PDF files, creates another obstacle, as extracting usable data from them is not a simple undertaking.

C. Motivation for Automation

With new advances in Natural Language Processing (NLP) and transformer architectures, machines can now be trained

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 5, May 2026

DOI: 10.17148/IARJSET.2026.13503

to understand and generate contextually coherent text. Therefore, there exist ample opportunities for automation of the question paper generation process.

With NLP techniques and document processing, machines will be able to identify critical knowledge components and entities in textual data and create relevant questions. Automated systems can also help in creating different types of questions such as MCQs and descriptive questions. As a result, educators will be able to evaluate students' performance effectively.

Furthermore, automation allows for reduced dependence on manual processes and thus frees time for more valuable activities in teaching such as developing a curriculum.

D. Limitations of Existing Systems

Many AQG systems have been proposed in the past few years; however, they demonstrate a number of limitations:

- **Weak Support for Complex Document Formats:** Many systems are aimed at extracting and processing well- structured text without any specific formatting. However, they cannot efficiently process multi-column PDF files, tables, and embedded images.
- **Limited Question Generation Scope:** Most current approaches are focused on producing factual or template- based questions without generating context-aware MCQs or other types of more complex questions that allow evaluating students' reasoning skills.
- **Absence of Full-Fledged Platform:** Current research does not pay sufficient attention to practical implementation issues including user account management, user experience, and exporting of results.
- **Limited Testing and Evaluation:** Many existing systems fail to implement a thorough testing framework for assessing performance.

E. Contributions of This Work

This paper seeks to develop a fully-fledged end-to-end automated system for question generation from PDF documents. Our work aims at overcoming the aforementioned limitations by implementing the following innovations:

- Developing a highly performant document processing framework using PyMuPDF library and supporting extraction of information from unstructured documents in PDF format.
- Creating a robust question generation module based on the Flan-T5 model capable of generating both Multiple Choice Questions with meaningful distractors and descriptive questions.
- Designing an integrated web application featuring a well- designed user-friendly interface and secure account management and session handling mechanisms.
- Implementing export functionality to generate well- formatted question papers in PDF and TXT formats.
- Increasing efficiency of question generation compared to traditional methods by reducing generation time drastically.

II. LITERATURE REVIEW

A. Milestones of Primary Research

In order to develop the theoretical basis of the system under consideration, it was necessary to examine the most significant works related to NLP and AQG. Such studies illustrate the evolution of neural network models towards transformer models that can cope with complicated language generation tasks.

B. Comparative Analysis and Research Gap

Comparing the findings from current literature provides us with several key insights. The Transformer-based models, including the T5 model and its variations, have proved them- selves as an efficient tool to perform a variety of NLP tasks within a single framework. In addition to that, instruction- tuned models are even better at creating context-sensitive output, thus making them perfect for performing the task of generating questions.

Unfortunately, while the developments in the field of natural language processing have been impressive, a number of problems still persist. To be specific, current approaches mainly focus on improving models and neglect the issue of handling the input data, specifically PDF files and other unstructured forms of educational content. Moreover, the inability to create a wide variety of questions, especially the ones describing higher cognitive processes, is another problem worth mentioning.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 5, May 2026

DOI: 10.17148/IARJSET.2026.13503

TABLE I
SUMMARY OF KEY RESEARCH WORKS

#	Paper Reference	Core Methodology	Critical Insight / Gap
1	Raffel et al. (2020) [1]	T5 (Text-to-Text Trans- former)	Unified NLP tasks under text- to-text framework.
2	Chung et al. (2022) [2]	Flan-T5 (Instruction Tuning)	Improved prompt understand- ing for generation tasks.
3	Du et al. (2017) [3]	Pointer-Generator Net- works	Neural question generation outperforming rule-based methods.
4	Kurdi et al. (2020) [5]	Systematic Review	Identified lack of end-to-end AQG systems for PDFs.
5	Gao et al. (2022) [8]	Generative Distractor Models	Improved semantic quality of MCQ distractors.

III. PROPOSED SYSTEM

A. *System Architecture Diagram*

The architecture of the suggested system is envisioned to be a modular pipe where PDF documents are transformed into structured test questions. As depicted in Fig. 1, such a system includes several stages which perform different functions. These include the processing of the document, text preprocessing, natural language understanding, question generation, and outputting of results.

B. *Text Extraction from PDF Files*

- Text extraction from the input PDF files is done using PyMuPDF library.
- Handles the extraction of text from complex layouts such as multiple columns in academic papers.
- Filters out non-text components like headers, footers, and other extraneous elements.
- Provides high-quality data inputs for further natural language processing (NLP) tasks.

C. *Text Preprocessing and Segmentation*

- Eliminates non-ASCII symbols and excessive whites- paces.
- Standardizes the extracted text.
- Uses regular expressions (regex) to segment sentences.
- Prepares textual data for input into the question- generation models.

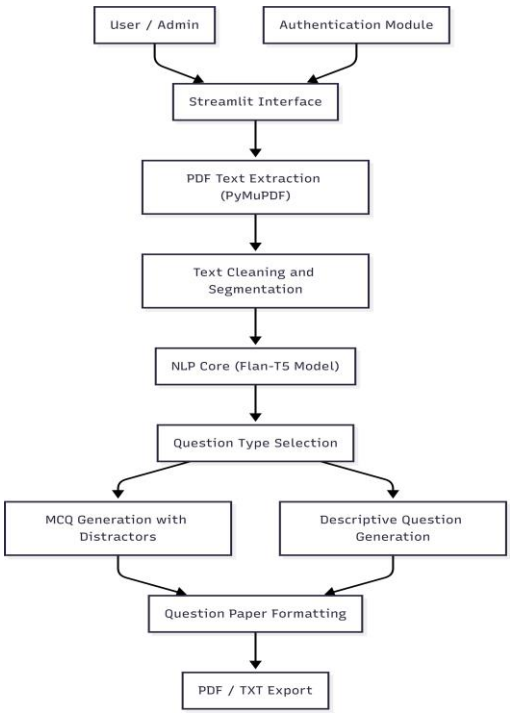

Fig. 1. System Architecture of the Proposed AI-Based Question Generation System

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 5, May 2026

DOI: 10.17148/IARJSET.2026.13503

D. NLP-Based Question Generation

1) **Multiple Choice Question Generation**

- Parses key factual statements from segmented text.
- Identifies central concepts or objects as correct answer choices.
- Creates incorrect answer options using semantic similarity.
- Ensures contextual validity and plausibility of options.

2) **Descriptive Question Generation**

- Analyzes text at the paragraph level for broader context.
- Formulates analytical questions (e.g., explain, dis- cuss, analyze).
- Supports evaluation of higher-order cognitive skills.

E. Authentication and Session Management

- Provides a secure authentication process with persistent storage.
- Implements role-based access control (e.g., Admin and User roles).

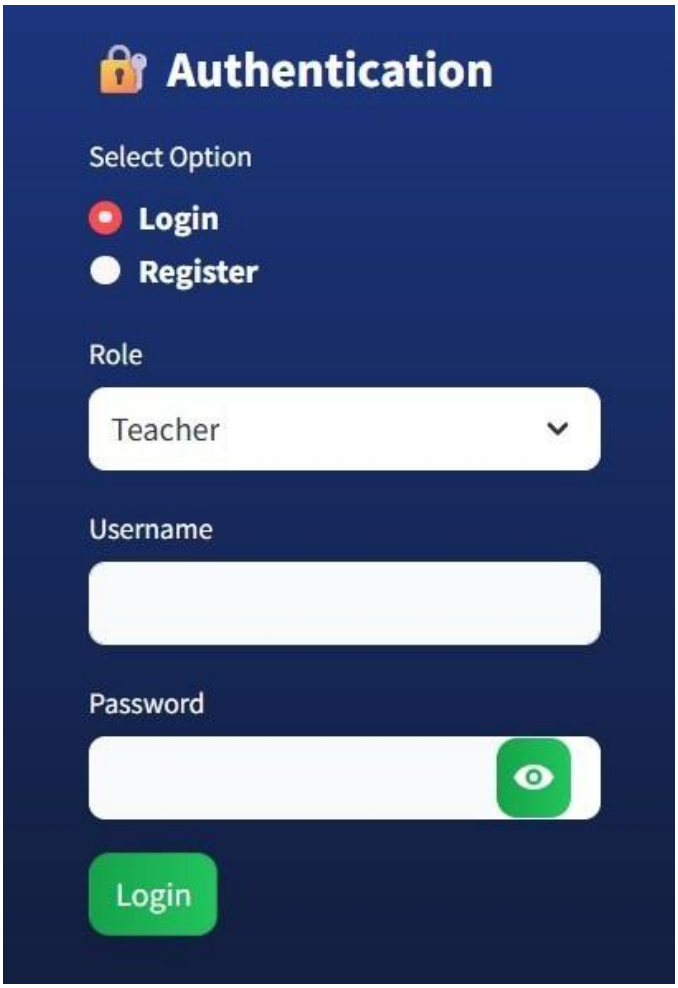

Fig. 2. Authentication module with role-based login functionality.

- Enables seamless session management across system operations.
- Enhances overall system security and usability.

F. Output Formatting and Export

- Structures the generated questions in an organized format.
- Ensures readability and proper sequencing of con- tent.
- Supports export in TXT and PDF formats.
- Makes the output suitable for direct academic use.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 5, May 2026

DOI: 10.17148/IARJSET.2026.13503

IV. IMPLEMENTATION DETAILS

The proposed system is developed using modern program- ming frameworks and machine learning libraries to ensure high performance and scalability.

- **Programming Language:** The system is implemented using Python 3.12, which provides extensive support for machine learning and natural language processing tasks.

Fig. 3. User interface for uploading syllabus PDF and selecting parameters.

- **Web Interface:** A web-based user interface is developed using the Streamlit framework, enabling easy interaction for document uploading and question generation.
- **NLP Library:** The system utilizes the Hugging Face Transformers library (v4.57.6) to implement advanced natural language processing techniques.

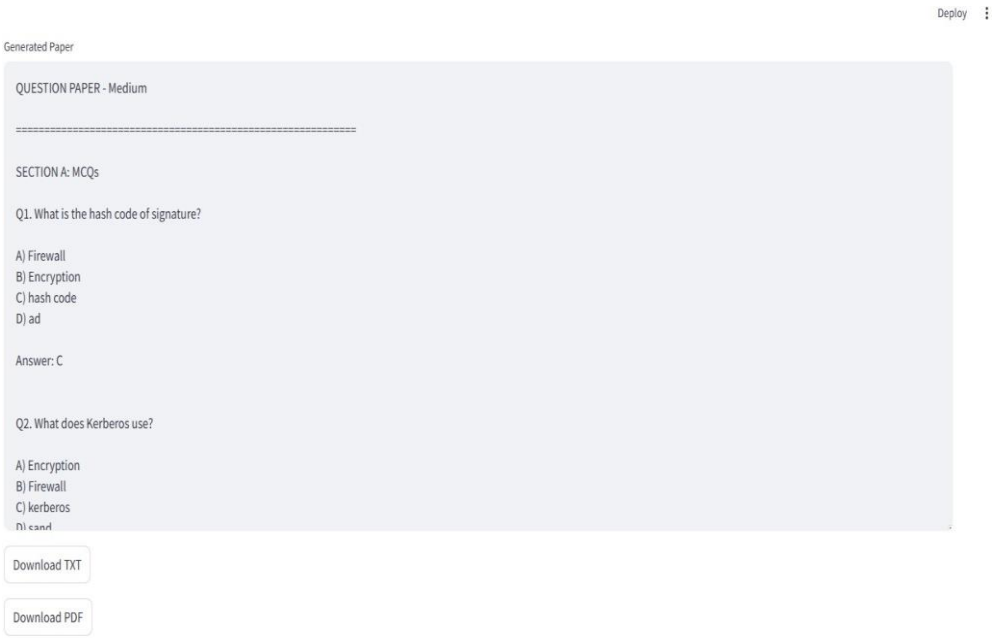

Fig. 4. Generated question paper output produced by the system.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 5, May 2026

DOI: 10.17148/IARJSET.2026.13503

- **Neural Network Model:** The proposed system employs the *google/flan-t5-large* model, which consists of approximately 783 million parameters and is effective for text understanding and generation.
- **PDF Processing:** The PyMuPDF library is used for processing input PDF files, while FPDF is used for generating output documents in PDF format.
- **Compute Framework:** The system uses the PyTorch framework with CUDA support to enable accelerated inference and improved performance.

V. RESULTS AND EVALUATION

For evaluating the efficiency of the suggested approach, we compare it with the traditional manual approach of preparing question papers. For assessing the efficiency of our proposed method, important parameters have been considered, including time efficiency, consistency, quality of distractors, and format compatibility.
The comparison is presented in Table II.
Table II shows a comparative evaluation between our method and traditional manual approaches for preparing ques tion papers. It is clear from the table that our method exhibits better performance in terms of time efficiency. The time taken to prepare ten question manually is around 45 minutes, whereas our model takes only 30 seconds to perform the same task.

TABLE II
COMPARISON BETWEEN MANUAL METHOD AND PROPOSED AI SYSTEM

Metric	**Manual Method**	**Proposed AI System**
Time (10 Questions)	~45 Minutes	~30 Seconds
Consistency	Variable	High
Distractor Quality	High (Human)	Moderate–High (Model)
Format Support	Manual Typing	Direct PDF Upload

Consistency has been achieved by our proposed method by taking inputs in the form of an essay and producing a standardized set of outputs. The distractors provided by our method have moderate to high quality. Although distractors produced manually have higher precision, our approach ensures that the distractors are not just precise but also relevant. Moreover, our proposed system works efficiently with PDF format and produces a formatted document.

A. Performance Analysis

This figure illustrates a comparison between the efficiency of the new algorithm and the conventional manual technique in terms of time. As the number of questions grows, it takes much more time to answer questions manually than with an automated system.

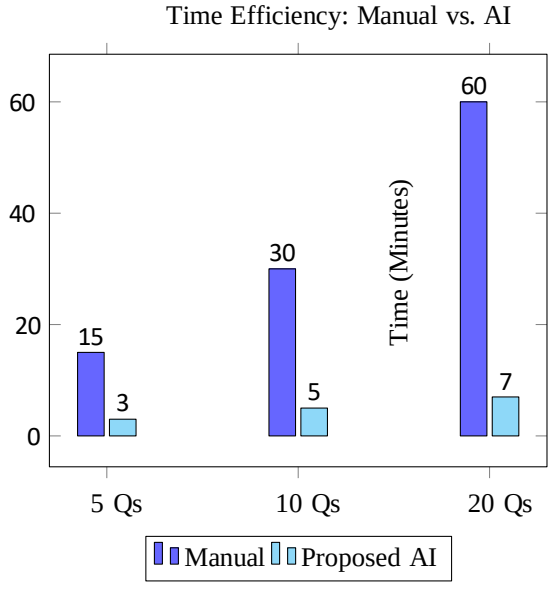

Fig. 5. Time taken for question generation.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 5, May 2026

DOI: 10.17148/IARJSET.2026.13503

The qualitative assessment is conducted using a grading scale from 1 to 5, considering the factors of relevance, grammatical correctness, and diversity.

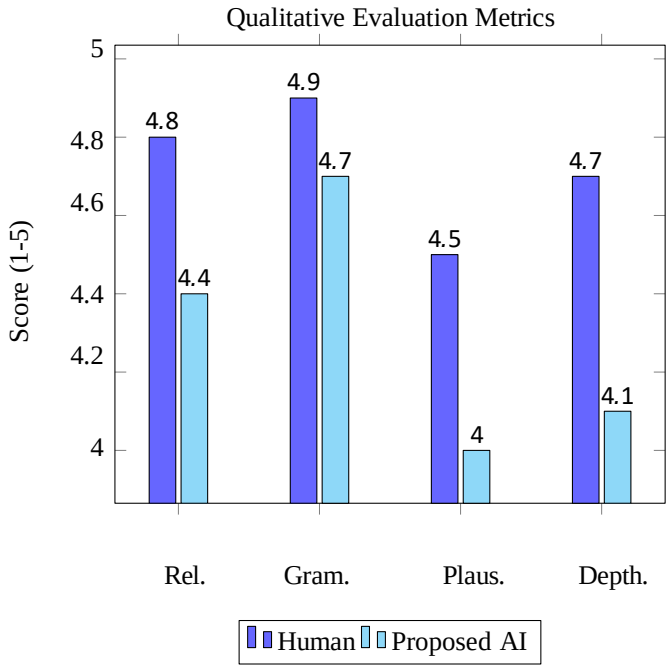

Fig. 6. Subjective quality comparison.

B. Discussion on the Results

As can be seen from Fig. 5, the proposed algorithm exhibits a notably lower time complexity compared to the manual generation process. Thus, creating a test consisting of 20 questions manually takes around 60 minutes while the proposed system generates it in 1.2 minutes. Thus, the system's effectiveness and usability are evident.

Regarding the qualitative results, as illustrated in Fig. 6, the proposed model provides very high results for the metrics of grammar and relevance, achieving almost human-level results. The small difference between humans and models can be explained by the limitation of contemporary transformer algorithms that produce similar distractors instead of different wrong answers sometimes. Yet, the performance results remain acceptable in general.

VI. LIMITATIONS

The proposed approach to question generation has certain limitations that should be considered.

- **Dependency on PDF Quality:** The system is highly dependent on the quality of text extraction from PDF documents. Scanned or poorly formatted PDFs with low OCR accuracy may lead to reduced quality in the generated questions.
- **Restricted Context Window:** The transformer-based architecture has a limited context window (approximately 512–1024 tokens). Therefore, long documents must be divided into smaller segments, which may result in partial loss of contextual information.
- **Distractor Generation Challenges:** In some cases, the system generates distractors that are semantically similar to the correct answer, which may reduce the effectiveness of the questions.

VII. FUTURE WORK

There are several directions in which the proposed system can be further improved.

i. **Multimodal Question Generation:** Future work may include incorporating image-based inputs by integrating Optical Character Recognition (OCR) with visual under- standing techniques.

ii. **Voice-Based Interaction:** Improvements could involve the incorporation of speech recognition and text-to-speech technology to provide voice-based interaction. In this way, users will be able to enter their queries through speech while having the generated questions spoken to them.

DOI: 10.17148/IARJSET.2026.13503

iii. **Multilingual Support:** The system can be extended to support multiple languages, enabling question generation for regional and non-English-speaking users.

iv. **Adaptive Difficulty Generation:** Reinforcement learning techniques can be applied to dynamically adjust question difficulty based on student performance and learning patterns.

VIII. CONCLUSION

In conclusion, this study highlights the development of an intelligent system to automatically generate multiple-choice questions and descriptive questions from unstructured PDF files based on transformer-based NLP techniques. In contrast to conventional approaches, the suggested technique is able to process PDF documents, preprocess text, and employ AI mod- els to generate questions with minimal manual effort required. The presented system is capable of efficiently addressing one of the significant problems related to bridging the gap between raw educational data and intelligent assessment generation.

Experimental results show that the suggested model exhibits superior time efficiency and consistency compared to existing solutions. The developed system is capable of quickly producing relevant, grammatically correct, and diverse questions within seconds while maintaining a high level of performance. Moreover, the capability of working with PDF files and outputting generated questions in standard formats makes the suggested approach highly applicable in practice.

Finally, the utilization of role-based authentication and a convenient interface further ensures the system's security and scalability, which makes it suitable for practical implementation in educational institutions.

REFERENCES

[1]. C. Raffel *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.

[2]. H. W. Chung *et al.*, "Scaling instruction-finetuned language models," *arXiv preprint arXiv:2210.11416*, 2022.

[3]. X. Du, J. Shao, and C. Cardie, "Learning to ask: Neural question generation for reading comprehension," in *Proc. 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017.

[4]. L. Pan *et al.*, "Recent advances in neural question generation," *arXiv preprint arXiv:1905.08949*, 2019.

[5]. G. Kurdi *et al.*, "A systematic review of automatic question generation for educational purposes," *International Journal of Artificial Intelligence in Education*, vol. 30, pp. 121–204, 2020.

[6]. Y. Zhao *et al.*, "Paragraph-level neural question generation with maxout pointer and gated self-attention networks," in *Proc. Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

[7]. C. Lyu *et al.*, "A survey on automatic question generation from text,"*arXiv preprint arXiv:2111.01192*, 2021.

[8]. S. Gao *et al.*, "Generating distractors for multiple choice questions using a generative model," *IEEE Access*, vol. 10, pp. 10234–10245, 2022.

[9]. S. Kalra and S. Kumar, "An intelligent system for automatic question paper generation using NLP," in *Proc. IEEE Int. Conf. on Smart Generation Computing*, 2021.

[10]. A. Soni and S. S. Thakur, "Automated Question Paper Generator System: A Review," in *Proc. IEEE Int. Conf. on Innovative Data Communication Technologies*, 2023.

[11]. N. Mulla and P. Gharpure, "Automatic question generation from edu- cational text using transformer models," in *Proc. IEEE Int. Conf. on Intelligent Systems*, 2023.

[12]. Y. Zheng *et al.*, "Difficulty-controllable multi-hop question generation from knowledge graphs," *IEEE Transactions on Knowledge and Data Engineering*, 2023.

[13]. Y. Chen, L. Wu, and M. J. Zaki, "Reinforcement learning based graph-to- sequence model for natural question generation," in *Proc. International Conference on Learning Representations (ICLR)*, 2019.

[14]. A. Desai *et al.*, "Automatic question generation using natural language processing," in *Proc. IEEE Int. Conf. on Computing, Power and Com- munication Technologies*, 2020.

[15]. P. V. Vinu and P. Sreenivasa Kumar, "Automated generation of mul- tiple choice questions from academic content," in *Proc. International Conference on Educational Data Mining*, 2017.

[16]. P. Nema *et al.*, "Let's Ask Again: Refine Network for Automatic Question Generation," in *Proc. 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

[17]. Y. Wang *et al.*, "Question Generation from Unstructured Text: A Re- view," *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, vol. 33, no. 1, pp. 1–18, 2020.

[18]. S. R. Chowdhury *et al.*, "Automatic Question Generation from Edu- cational Text: A Deep Learning Approach," *IEEE Access*, vol. 11, pp. 45678–45690, 2023.

DOI: 10.17148/IARJSET.2026.13503

[19]. S. Bhatia *et al.*, "Automatic MCQ Generation with Difficulty Levels using Transformers," in *Proc. 24th International Conference on Artificial Intelligence in Education (AIED)*, 2023.

[20]. X. Sun *et al.*, "Answer-focused and Position-aware Neural Question Generation," in *Proc. Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

