

A Comprehensive Study on NLP Models for Automatic Question Generation

Sheetal A Rawoot, Pro. Dr. Mr. B. D. Phulpagar

Department of Computer Engineering, P.E.S Modern College of Engineering, Pune, Maharashtra, India

Department of Computer Engineering, P.E.S Modern College of Engineering, Pune, Maharashtra, India

ABSTRACT: Automatic Question Generation (AQG) is an important area of Natural Language Processing (NLP) and Artificial Intelligence (AI). AQG systems help computers automatically generate questions from textual materials for educational and assessment purposes. These systems are widely used in online education platforms, intelligent tutoring systems, e-learning applications, and educational chatbots. Earlier AQG systems mainly relied on rule-based and template-based approaches that depended on predefined grammar rules and sentence structures [5], [6]. Although these methods generated simple questions, they lacked flexibility, semantic understanding, and contextual awareness.

With the advancement of Deep Learning and Transformer-based architectures such as BERT, GPT, T5, BART, and PEGASUS, AQG systems have significantly improved in terms of contextual understanding, semantic reasoning, and fluency [1], [2], [9], [14]. These models utilize self-attention mechanisms to capture semantic relationships and long-range dependencies in text, enabling the generation of human-like and meaningful questions. Modern AQG systems also integrate semantic similarity techniques such as Sentence-BERT (SBERT) for automatic answer evaluation by comparing semantic meaning instead of exact keyword matching [3].

This paper presents a comprehensive study of various AQG methodologies including rule-based systems, machine learning techniques, neural network models, Transformer-based architectures, and Large Language Models. The study also discusses datasets, evaluation metrics such as BLEU and ROUGE, semantic similarity techniques, experimental results, challenges, and future research directions in AQG systems [7], [8], [15].

KEYWORDS: Natural Language Processing, Automatic Question Generation, Transformers, Deep Learning, Semantic Similarity

I. INTRODUCTION

Natural Language Processing (NLP) is a branch of Artificial Intelligence (AI) that enables computers to understand, process, and generate human language. One of the important applications of NLP in education is Automatic Question Generation (AQG). AQG systems automatically create questions from textual content such as study materials, articles, books, and documents. Nowadays, online learning and digital education platforms are growing rapidly, increasing the demand for intelligent educational technologies. Teachers spend significant time preparing examination papers, quizzes, and assessment materials manually. AQG systems help reduce this effort by automatically generating relevant and meaningful questions while maintaining educational quality and contextual accuracy [6], [7].

Initially, AQG systems mainly used rule-based and template-based approaches that depended on predefined grammar rules, sentence structures, and linguistic patterns [5], [6]. These systems transformed declarative sentences into interrogative forms using syntactic rules and keyword identification techniques. Although these methods were simple and computationally efficient, they lacked flexibility, semantic understanding, and contextual awareness. As a result, they struggled to generate complex reasoning-based and context-sensitive educational questions [7].

To improve AQG performance, researchers introduced machine learning and statistical approaches that learned language patterns directly from datasets. Neural Sequence-to-Sequence (Seq2Seq) models with attention mechanisms further improved grammatical correctness and fluency in generated questions [5], [8]. However, recurrent neural networks still faced difficulties in understanding long-range contextual dependencies and semantic relationships within lengthy textual passages.

Volume 15, Issue 5, May 2026**|DOI: 10.15680/IJRSET.2026.1505140|**

A major advancement in AQG occurred with the introduction of Deep Learning and Transformer architectures. Models such as GPT, BERT, T5, BART, and PEGASUS significantly improved AQG systems by using self-attention mechanisms to capture semantic meaning, contextual relationships, and long-range dependencies in text [1], [2], [9], [14]. These Transformer-based models generate fluent, context-aware, and human-like questions with improved semantic accuracy and reasoning capability.

Modern AQG systems are also integrated with automatic answer evaluation techniques using semantic similarity methods such as Sentence-BERT (SBERT) and cosine similarity [3]. These techniques evaluate student answers based on semantic meaning instead of exact keyword matching, improving fairness and accuracy in educational assessment systems.

Recent AQG research focuses on advanced methodologies including multi-hop reasoning, graph attention networks, reinforcement learning, dependency-guided attention mechanisms, and prompt-based Large Language Models [4], [11], [12], [15]. These approaches enhance contextual understanding, question diversity, and reasoning capability in educational applications. Transformer and LLM-based AQG systems are now widely applied in intelligent tutoring systems, educational chatbots, automated examinations, and adaptive learning environments.

This paper presents a comprehensive study of AQG techniques, methodologies, datasets, evaluation metrics, experimental results, challenges, and future research directions in educational NLP applications. The study analyzes the evolution of AQG systems from rule-based approaches to advanced Transformer and Large Language Model architectures and highlights their impact on modern intelligent educational technologies [1]–[15].

II. LITERATURE SURVEY

The field of Automated Question Generation (AQG) has evolved significantly from simple rule-based systems to advanced neural and Transformer-based architectures capable of generating fluent, contextually relevant, and semantically meaningful questions. Researchers have continuously improved AQG methodologies to enhance contextual understanding, reasoning capability, grammatical correctness, and educational applicability. This literature survey examines the evolution of AQG methodologies, specialized architectures, domain-specific applications, prompt-based learning techniques, evaluation metrics, and existing research challenges [6], [7], [8].

1. Evolution of Methodological Frameworks**a) Rule-Based and Heuristic Methods**

Early AQG systems mainly relied on hand-crafted templates, grammar rules, syntactic parsing, and linguistic features such as semantic role labeling to transform declarative sentences into interrogative forms [5], [6]. These systems were simple and domain-independent but required extensive manual rule creation and lacked flexibility in handling diverse sentence structures and semantic variations. As a result, they struggled to generate complex and context-aware educational questions.

b) Neural Sequence-to-Sequence (Seq2Seq) Models

With the advancement of Deep Learning, AQG systems transitioned toward neural Sequence-to-Sequence (Seq2Seq) architectures using Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks [5], [8]. Attention mechanisms were introduced to improve contextual focus during decoding. These models generated more fluent and grammatically correct questions compared with rule-based systems. However, the sequential nature of RNNs limited their ability to capture long-range contextual dependencies, especially for paragraph-level question generation.

c) The Transformer Revolution

The introduction of Transformer architectures fundamentally transformed AQG research through the use of self-attention mechanisms that capture global contextual dependencies within text [1], [2]. Models such as BERT, GPT, T5, and BART significantly improved semantic understanding, contextual reasoning, and generation fluency. Transformer-based AQG systems outperform earlier neural architectures in BLEU, ROUGE, and human evaluation metrics [1], [9], [14].

2. Advanced Specialized Architectures**a) Sequential BERT (BERT-HLSQG)**

Researchers developed the BERT-HLSQG architecture to overcome the limitations of standard BERT in generative tasks [14]. This methodology restructures BERT into a sequential generation framework that incorporates information

from previously decoded outputs. Highlight tokens [HL] are used to mark answer spans within textual passages, reducing ambiguity and improving paragraph-level contextual understanding.

b) Knowledge Graph Integration (KG4QG)

For complex reasoning-based AQG, the KG4QG framework generates knowledge graphs from textual passages to establish entity-relation connections [4]. Graph Attention Networks (GAT) are then applied to learn semantic relationships among interconnected entities. This methodology improves factual consistency and multi-hop reasoning capability across multiple paragraphs.

c) Answer-Agnostic Systems (QGAE)

Traditional AQG systems are generally answer-aware and require predefined answer spans. In contrast, the QGAE framework provides an end-to-end solution for generating question-answer pairs directly from raw text [11]. It employs dual encoders for answer extraction and question generation simultaneously, making the system more efficient and portable compared with multi-stage AQG pipelines.

3. Domain Adaptation and Specialized Datasets

a) Educational Domain

Transformer-based models such as T5 have shown excellent performance in generating factual, inferential, and multiple-choice questions from educational materials [9], [10]. Specialized educational datasets such as RACE4QG and EduProbe were developed using NCERT-based educational content to train models for reasoning-oriented and comprehension-based question generation rather than simple factual recall.

b) Ancient Scriptures and Technical Manuals

AQG systems have also been applied to specialized domains such as ancient scriptures and technical documentation [18]. The Vadic-BAG framework combines BERT embeddings, GRU networks, and attention mechanisms for generating questions from Atharv Ved texts. Similarly, Semantic Role Labeling (SRL)-based AQG systems improve generalization when adapting models from Wikipedia datasets to technical manuals and domain-specific content.

c) Cross-Domain Generalization

Cross-domain AQG remains an important research challenge. SRL-Seq2Seq systems utilize generalized semantic representations to improve adaptability across multiple domains [7]. These systems achieve performance comparable to large-scale Large Language Models while using smaller architectures and reduced computational resources.

4. Prompt-Based Techniques and Large Language Models

a) LLM Performance

Recent AQG research investigates the use of Large Language Models such as GPT-3.5-Turbo and Davinci for question generation without extensive task-specific fine-tuning [15]. While Transformer models such as T5 often achieve higher automated evaluation scores, human evaluators frequently rate LLM-generated questions higher in terms of novelty, diversity, and complexity.

The Power of Prompts:

Prompt engineering significantly influences the quality of generated questions in LLM-based AQG systems. Long prompts containing detailed contextual cues generate more comprehensive and reasoning-based questions, while shorter prompts produce concise and focused outputs [15]. Prompt-based AQG enables flexible and adaptive educational content generation.

Comparison of AQG Approaches Table 1: Comparison of AQG Techniques

Approach	Advantages	Disadvantages	Performance
Rule-Based	Simple	Not flexible	Low
ML-Based	Better than rules	Weak context	Medium
Seq2Seq	Fluent output	Limited context	Medium
Attention	Better understanding	Moderate complexity	High
Transformer	Accurate	High cost	Very High

III. METHODOLOGY

The methodology of Automatic Question Generation (AQG) systems has evolved from traditional rule-based approaches to advanced Transformer and Large Language Model (LLM)-based architectures. Modern AQG methodologies focus on generating context-aware, semantically meaningful, and grammatically correct questions from

textual data. This section explains the major AQG methodologies, specialized architectures, preprocessing techniques, and evaluation approaches used in recent research [6], [7].

1. Rule-Based Methods

Initially, AQG systems mainly relied on grammar rules, templates, syntactic parsing, and pattern-matching techniques to convert declarative sentences into interrogative forms [5], [6]. These systems used manually designed linguistic rules such as Part-of-Speech (POS) tagging, Named Entity Recognition (NER), and dependency parsing for identifying sentence structure and important keywords.

Although rule-based AQG systems were simple and computationally efficient, they lacked flexibility and semantic understanding. They struggled to handle complex sentence structures, contextual reasoning, and higher-order educational questions [7].

2. Transformer-Based Fine-Tuning

Modern AQG systems commonly use large-scale pre-trained Transformer models that are fine-tuned on task-specific educational datasets [1], [2]. Transformer architectures use self-attention mechanisms to capture semantic relationships and contextual dependencies within text.

a) Unified Text-to-Text Framework (T5)

T5 treats every NLP task as a text-to-text generation problem [9], [10]. Researchers fine-tuned T5 on educational datasets to generate factual, inferential, and multiple-choice questions dynamically. T5 demonstrated superior contextual understanding and semantic fluency in educational AQG tasks. It also achieved excellent performance in abstractive summarization tasks.

b) Denoising and Gap-Sentence Generation (BART & PEGASUS)

BART uses a bidirectional encoder and autoregressive decoder architecture for text generation [1]. BART is trained using denoising objectives where corrupted input text is reconstructed during training, improving contextual learning and semantic understanding.

PEGASUS is specifically pre-trained using gap-sentence generation techniques, where important sentences are removed from a passage and reconstructed during training [2]. This methodology is highly effective for abstractive summarization and educational question generation tasks.

3. Specialized Question Generation Architectures

Several researchers proposed specialized modifications to standard Transformer architectures to address specific AQG challenges such as ambiguity reduction, multi-hop reasoning, and domain adaptation.

a) Sequential BERT (BERT-HLSQG)

BERT was restructured into a sequential generation framework called BERT-HLSQG to support paragraph-level AQG tasks [14]. This methodology uses Highlight Tokens [HL] to mark answer spans within passages, reducing ambiguity when the same answer appears multiple times.

b) End-to-End Answer-Agnostic QG (QGAE)

The QGAE framework performs answer extraction and question generation simultaneously using dual encoders and a shared decoder [11]. One encoder extracts candidate answer spans, while the second encoder reconstructs contextual information. This approach enables direct generation of question-answer pairs from raw textual content without requiring predefined answers.

c) Vadic-BAG Model

The Vadic-BAG framework was designed for question generation from ancient scripture texts such as Atharv Ved [18]. This methodology combines BERT embeddings, Gated Recurrent Units (GRU), and attention mechanisms to process context-rich and semantically complex textual data. Initially, the T5 model is used to generate question-answer pairs, which are further processed by the hybrid architecture.

4. Knowledge-Guided and Linguistic Methodologies

Modern AQG systems increasingly integrate external knowledge structures and linguistic representations to improve reasoning capability and semantic consistency.

a) Knowledge Graph Integration (KG4QG)

The KG4QG framework generates knowledge graphs from textual passages to identify entity-relation triplets [4]. Graph Attention Networks (GAT) are then applied to generate graph embeddings, which are combined with sentence embeddings to support multi-hop reasoning across interconnected paragraphs.

b) Answer-Oriented Graph Attention Networks (AO-GAT)

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505140|

AO-GAT constructs dependency graphs for passages and connects sentences containing answer words [11]. A Graph Attention Network updates word embeddings based on neighboring attention weights, allowing the model to capture deep intra-sentence and inter-sentence semantic dependencies.

c)SRL-Seq2Seq

Semantic Role Labeling (SRL)-based Seq2Seq models convert textual sentences into generalized semantic representations [7]. Training on semantic role labels instead of raw text reduces domain dependency and improves cross-domain adaptability when shifting from general datasets such as Wikipedia to technical manuals and specialized educational content.

5. Prompt-Based Methodologies

Recent AQG research investigates the influence of prompt engineering on question generation quality in Large Language Models.

Long vs. Short Prompts

Researchers evaluated different prompting strategies including:

- a) Long prompts with detailed contextual cues
- b) Short prompts with condensed key information
- c) Instruction-only prompts without contextual guidance

IV. EXPERIMENTAL RESULTS

Experimental results across the sources demonstrate that Transformer-based models consistently outperform traditional architectures in accuracy, fluency, and reasoning capabilities [1], [9], [15].

1. Automated Question Generation (AQG) Results

SQuAD Benchmark (Paragraph Level): The BERT-HLSQG model advanced the state-of-the-art BLEU-4 score to 22.17, surpassing earlier RNN-based models which achieved 16.85 [14].

Multi-hop Reasoning (HotpotQA): The KG4QG framework based on BART achieved a BLEU-4 score of 38.56 and a ROUGE-L score of 47.11, significantly outperforming the MulQG baseline [4].

Educational Accuracy: A T5-based model for school assessments reached an average precision of 89.5% and an accuracy of 91.2% on academic texts [9], [10].

Ancient Scriptures: The Vadic-BAG model achieved a 94% accuracy rate on Atharv Ved datasets and 95% on SQuAD datasets, proving effective for complex and archaic textual content [18].

Answer-Agnostic Generation: The QGAE model achieved an Exact Match (EM) score of 53.82% for answer extraction and a BLEU-4 score of 20.11 for question generation, nearly doubling the performance of conventional multi-stage baselines [11].

2. Text Summarisation Results

News Articles (CNN/DailyMail): The T5 Transformer achieved the highest scores across all evaluation metrics, including ROUGE-1 of 35.12 and ROUGE-L of 32.82. The model outperformed GPT, BART, and PEGASUS [1], [2], [9].

Non-Mainstream Languages (Czech): The CzeGPT-2 model with 117M parameters achieved a ROUGE-L score of 12.5, comparable to the much larger mBART model which achieved 13.5, showing that smaller language-specific models can achieve efficient performance [16].

3. Cross-Domain and Human Evaluation

Cross-Domain Generalisation: The SRL-Seq2Seq model based on BART achieved a BLEU-4 score of 20.2 on technical car manual datasets, comparable to the larger GPT-4o model which achieved 19.3 [7].

Human Qualitative Assessment: In school-level question generation tasks, human evaluators rated Text-Davinci-003 higher than human-generated baselines for novelty with a score of 3.73 compared to 3.10 and complexity with a score of 4.25 compared to 4.13 [15].

Efficiency: The proposed T5-based AQG model achieved an inference time of approximately 0.8 seconds per generated question on high-end hardware, making it suitable for real-time educational platforms and intelligent tutoring systems [9].

The experimental analysis demonstrates that:

Transformer architectures significantly improve AQG quality and contextual understanding.

T5, BART, and PEGASUS achieve superior BLEU and ROUGE performance.

Graph attention and multi-hop reasoning frameworks improve semantic reasoning capability.
Hybrid architectures effectively process domain-specific and complex textual data.
Prompt-based Large Language Models generate highly novel and semantically rich questions.
Lightweight Transformer models provide efficient performance with reduced computational requirements.

V. CONCLUSION

Automatic Question Generation (AQG) has evolved significantly from traditional rule-based and template-based systems to advanced Transformer and Large Language Model (LLM) architectures. Earlier AQG systems mainly relied on grammar rules and syntactic patterns, which limited their flexibility, contextual understanding, and semantic reasoning capabilities [5], [6]. With the advancement of Deep Learning and self-attention mechanisms, modern AQG systems can now generate fluent, meaningful, and context-aware questions from complex textual content.

Transformer-based models such as BERT, GPT, T5, BART, and PEGASUS have significantly improved AQG performance in educational applications by enhancing contextual understanding, semantic accuracy, reasoning capability, and generation fluency [1], [2], [9], [14]. These models effectively capture long-range dependencies and semantic relationships in textual data, enabling the generation of high-quality educational questions.

Semantic similarity techniques such as Sentence-BERT (SBERT) have further improved automatic answer evaluation systems by assessing semantic meaning instead of exact keyword matching [3]. Advanced methodologies including graph attention networks, multi-hop reasoning, reinforcement learning, prompt engineering, and knowledge graph integration have enhanced the capability of AQG systems to generate more diverse, contextually rich, and reasoning-oriented questions [4], [11], [12], [15].

Experimental results demonstrate that Transformer and LLM-based AQG systems consistently outperform traditional rule-based and neural architectures in BLEU, ROUGE, semantic similarity, and human evaluation metrics [9], [11], [15]. Specialized AQG frameworks have also shown promising results in multilingual learning systems, educational assessments, technical documentation, and domain-specific applications such as ancient scripture analysis [16], [18].

Despite these advancements, AQG systems still face several challenges including high computational cost, explainability issues, multilingual adaptation, bias in training data, ethical concerns, and difficulty generating higher-order reasoning questions [7], [15]. Ongoing research is continuously improving AQG methodologies to make educational AI systems more efficient, interpretable, and accessible.

In the future, AQG systems are expected to support multilingual, adaptive, personalized, and domain-specific educational environments. Integration with intelligent tutoring systems, conversational AI, and real-time assessment platforms will further enhance digital learning technologies and contribute significantly to the development of intelligent educational ecosystems.

REFERENCES

- 1] M. Lewis et al., "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation," in Proceedings of ACL, pp. 7871–7880, 2020.
- 2] J. Zhang et al., "PEGASUS: Pre-training with Extracted Gap-Sentences for Abstractive Summarization," in Proceedings of ICML, pp. 11328–11339, 2020.
- 3] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in Proceedings of EMNLP, pp. 3982–3992, 2019.
- 4] Efficient Multi-hop Question Generation Author links open overlay panel John Emerson, Yllias Chali 2023
- 5] Q. Zhou et al., "Neural Question Generation from Text: A Preliminary Study," arXiv preprint arXiv:1704.01792, 2017. arXiv
- 6] S. Kurdi et al., "A Systematic Review of Automatic Question Generation for Educational Purposes," Educational Technology & Society, vol. 23, no. 1, pp. 1–16, 2020.
- 7] N. Mulla and P. Gharpure, "Automatic Question Generation: A Review of Methodologies, Datasets, Evaluation Metrics, and Applications," Progress in Artificial Intelligence, vol. 12, pp. 1–32, 2023. Springer

Volume 15, Issue 5, May 2026**|DOI: 10.15680/IJRSET.2026.1505140|**

- 8] B. Das et al., “Automatic Question Generation and Answer Assessment: A Survey,” Research and Practice in Technology Enhanced Learning, vol. 16, article 5, 2021. Springer
- 9] K. Grover et al., “Deep Learning Based Question Generation using T5 Transformer,” Recent Advances in Computer Science and Communications, 2021. ScienceDirect
- 10] Proadpran Punyabukkana “English Grammar Multiple-Choice Question Generation using Text-to-Text Transfer Transformer,” Computers and Education: Artificial Intelligence, vol. 4, p. 100158, 2023.
- 11] “RTRL: Relation-aware Transformer with Reinforcement Learning for Deep Question Generation,” Knowledge-Based Systems, vol. 300, p. 112120, 2024. ScienceDirect
- 12] J li “Syntax-aware Question Generation through Dependency Relations-guided Attention,” Engineering Applications of Artificial Intelligence, 2025. ScienceDirect
- 13] “Context-based Automatic Question Generation for Educational Assessment via Lightweight Transformer Augmentation,” Engineering Applications of Artificial Intelligence, 2026. ScienceDirect
- 14] Ying, Yao, A Recurrent BERT-based Model for Question Generation - ACL Anthology
- 15] MOHAIMENUL AZAM KHAN RAIAN 1, MD. SADDAM HOSSAIN MUKTA
A_Review_on_Large_Language_Models_Architectures_Applications_Taxonomies_Open_Issues_and_Challenges.pdf.
2024. IEEE
- 16] ADAM HÁJEK AND ALEŠ HORÁK, CzeGPT-2—Training New Model for Czech Generative Text Processing Evaluated With the Summarization Task ,2024
- 17] DAVID DE-FITERO-DOMINGUEZ, EVA GARCIA-LOPEZ , ANTONIO GARCIA-CABOT , JESUS-ANGEL DEL-HOYO-GABALDON, AND ANTONIO MORENO-CEDIEL
Distractor_Generation_Through_Text-to-Text_Transformer_Models.pdf, Departamento de Ciencias de la Computación, Universidad de Alcalá, Edificio Politécnico, Alcalá de Henares, 28871 Madrid, Spain, 2024
- 18] ABHISHEK KUMAR PANDEY AND SANJIBAN SEKHAR ROY , (Member, IEEE) Vellore Institute of Technology, Vellore, 632014, India, Extractive_Question_Answering_Over_Ancient_Scriptures_Texts_Using_Generative_AI_and_Natural_Language_Processing_Techniques.pdf, 2024



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

 9940 572 462  6381 907 438  ijirset@gmail.com



www.ijirset.com

Scan to save the contact details