A literature review of research on question generation in education

Xiaohui Dong, Xinyu Zhang, Zhengluo Li, Quanxin Hou, Jixiang Xue and Xiaoyi Li

College of Computer Science & Engineering, Northwest Normal University, Lanzhou, Gansu, China

ABSTRACT

As a key natural language processing (NLP) task, question generation (QG) is crucial for boosting educational quality and fostering personalized learning. This article offers an in-depth review of the research advancements and future directions in QG in education (QGEd). We start by tracing the evolution of QG and QGEd. Next, we explore the current state of QGEd research through three dimensions: its three core objectives, commonly used datasets, and question quality evaluation methods. This article also underscores its unique contributions to QGEd, including a systematic analysis of the research landscape and an identification of pivotal challenges and opportunities. Lastly, we highlight future research directions, emphasizing the need for deeper exploration in QGEd regarding multimodal data processing, controllability of fine-grained cognitive and difficulty levels, specialized educational dataset construction, automatic evaluation technology development, and system architecture design. Overall, this review aims to provide a comprehensive overview of the field, offering valuable insights for researchers and practitioners in educational technology.

Subjects Human-Computer Interaction, Artificial Intelligence, Computer Education, Natural Language and Speech
Keywords Question generation, Educational application, Natural language generation, Survey

Submitted 21 November 2024
Accepted 18 August 2025
Published 16 September 2025

Corresponding author
Xiaohui Dong, dxh@nwnu.edu.cn

Academic editor
Varun Gupta

Additional Information and Declarations can be found on page 30

DOI **10.7717/peerj-cs.3203**

INTRODUCTION

Formulating questions is an indispensable part of the educational process. Well-crafted questions can not only inspire students to grasp the content of teaching but also check students' understanding of knowledge, stimulate their interest in learning, and promote the development of their thinking. In traditional educational models, the design of questions is usually manually completed by subject teachers or educational experts. This method requires a significant amount of time and cognitive effort, and the quality of the questions is largely limited by the professional level and scope of attention of the designers. With the rapid growth of knowledge and information, this traditional manual questioning approach has become difficult to adapt to the needs of large-scale online learning and ubiquitous learning in the current digital age. Automatic question generation (AQG) technology in education can assist in developing teaching resources, promote personalized teaching, enhance educational quality, and support teachers' professional development. In summary, AQG has tremendous potential to empower education, and AQG in education holds significant research and application value.

How to cite this article Dong X, Zhang X, Li Z, Hou Q, Xue J, Li X. 2025. A literature review of research on question generation in education. **PeerJ Comput. Sci. 11:e3203 DOI 10.7717/peerj-cs.3203**

Table 1 **The instance from the XQuAD (*Artetxe, Ruder & Yogatama, 2020*) dataset suitable for answer-aware question generation.**

Input Paragraph: The heat required for boiling the water and supplying the steam can be derived from various sources, most commonly from **[burning combustible materials]**$_1$ with an appropriate supply of air in a closed space (called variously **[combustion chamber]**$_2$, firebox). In some cases, the heat source is a nuclear reactor, geothermal energy, **[solar]**$_3$ energy or waste heat from an internal combustion engine or industrial process. In the case of model or toy steam engines, the heat source can be an **[electric]**$_4$ heating element.
Answer Span: burning combustible materials, combustion chamber, solar, electric
Question1: What is the usual source of heat for boiling water in the steam engine? **Question2:** Aside from firebox, what is another name for the space in which combustible material is burned in the engine? **Question3:** Along with nuclear, geothermal and internal combustion engine waste heat, what sort of energy might supply the heat for a steam engine? **Question4:** What type of heating element is often used in toy steam engines?

AQG abbreviated as question generation (QG), has its research history that dates back to the 1970s (*Wolfe, 1976*). It is an important and challenging task in cognitive intelligence, aimed at enabling computers to generate content-relevant and fluently expressed questions from various inputs, such as natural language texts (*Zhang et al., 2022*), knowledge bases (*Reddy et al., 2017*; *Guo et al., 2023*; *Bi et al., 2024*), images (*Mulla & Gharpure, 2023*), and videos (*Priya et al., 2022*). This reflects the process of a machine moving from understanding knowledge, discovering knowledge, to applying knowledge. QG can generally be formally defined as: given a natural language text $D = \{x_1, \ldots, x_n\}$, the goal is to generate natural language questions $Q = \{y_1, \ldots, y_{|y|}\}$, that is:

$$\begin{aligned} \overline{Q} &= \arg\max P(Q|D) \\ &= \arg\max_{y} \sum_{i=1}^{|y|} \log P(y_i|D, y_{<i}) \end{aligned} \tag{1}$$

x_i denotes a word within the input text, n signifies the length of the input text, y_i refers to a word within the generated question, and $|y|$ indicates the sentence length of the generated question (*Zhang et al., 2022*).

There are multiple ways to classify QG tasks. Typically, these tasks are categorized based on whether the focus is on the answer to the question, that is, whether the input contains an answer, into answer-aware (answer-sensitive) and answer-unaware QG. In answer-aware QG, questions are formulated towards a given answer within the context, as shown in Table 1, where the input paragraph represents the provided context, and the answer span indicates the position of the target answer within it. These answers also serve as inputs for answer-aware QG. Question1 to question4 are the corresponding questions generated by the machine based on the provided context and the target answer. In contrast, answer-unaware QG is not constrained by a target answer and only requires the formulation of questions based on the given context information (document, paragraph, sentence, or keywords), as shown in Table 2, where the input document is the provided context, and question1 to question4 are the questions generated by the machine. The answers to these questions may either be present in the original text or simply related to its content.

Table 2 **The instance from the LearningQ (*Chen et al., 2018*) dataset suitable for answer-unaware question generation.**

Input Document: … Why is this region named this way? What is it in the middle of or near to? It is the proximity of these countries to the West (to Europe) that led this area to be termed "the near east." Ancient Near Eastern Art has long been part of the history of Western art, but history didn't have to be written this way. It is largely because of the West's interests in the Biblical "Holy Land" that ancient Near Eastern materials have been regarded as part of the Western canon of the history of art …
Question1: So, Mesopotamia is a name that could be used interchangeably with the Middle East or the Near East? **Question2:** How does it compare to a city like New York city? **Question3:** What were the conventions of art in the Ancient Middle East? **Question4:** What about the lost gardens of Babylon?

With the gradual maturation of general-purpose large language models (LLMs), some studies have begun to explore prompt engineering as a less costly and more efficient approach to guide these models in generating the desired questions. For instance, *Lee et al. (2024)* employed few-shot prompting to direct the model to generate questions of varying cognitive levels within a 2D matrix framework that encompasses multiple question types and formats. *Tonga, Clement & Oudeyer (2024)* designed error-based prompts and general prompts based on the principles of chain-of-thought prompting, enabling the "teacher LLM" to generate heuristic questions in response to the reactions of the "student LLM".

Rational

Regarding QG in education (QGEd), there are also some relevant reviews. We found that some reviews did not explicitly propose the characteristics and objectives of QGEd, or they were not sufficiently focused on the educational field. For instance, *Le, Kojiri & Pinkwart (2014)* reviewed the research progress of QGEd, categorizing it into three types from the perspective of application objectives: knowledge/skill acquisition, knowledge evaluation, and tutoring dialogues. *Kurdi et al. (2020)* also provided a review of this field, but there was insufficient discussion on the guidance and integration of educational theories in generating questions. In 2024, *Al Faraby, Adiwijaya & Romadhony (2024)* conducted a review of QGEd from 2016 to the beginning of 2022, focusing on analyzing the technical components of QG, such as context, answers, questions, and datasets. Although this study concentrated on reviewing methods based on neural network models and involved a relatively small number of literatures, it clearly proposed that naturalness, diversity, controllability, practicality (educational value), and personalization are the core objectives of research in this field, providing a clear direction for QGEd research:

(1) Naturalness means that the generated questions should not only conform to the expression habits and grammatical rules of the target language but also ensure that the questions can be answered from the context or at least can be resolved using background knowledge related to the context. This goal is a universal requirement for all QG tasks, not limited to QGEd.

(2) Diversity refers to generating a variety of different types of questions for the same context to avoid producing overly general and similar questions, in order to meet the needs of different teaching scenarios.

(3) Controllability mainly points to the cognitive level and difficulty level of the generated questions, that is, guiding the model to generate questions of different cognitive levels or difficulty levels through corresponding mechanisms.

(4) Educational value emphasizes that the generated questions should focus on teaching objectives and content, avoiding the generation of some popularized and trivial questions.

(5) Personalization aims to meet the individual differences of learners and specific educational needs.

Among these objectives, naturalness has been well addressed in previous research. For instance, in studies related to heuristic rule-based methods, language features are often utilized, such as named entity recognition (*Sang & Meulder, 2003*), positional tagging (*Lindberg et al., 2013*), or by using placeholders (*Jia et al., 2021*; *Kumar, Banchs & D'Haro, 2015*; *Olney, Pavlik & Maass, 2017*) to replace keywords or sentences in the context; in neural question generation (NQG) methods, many works have improved the quality of questions by refining the encoding and decoding processes (*Wang et al., 2019*; *Song et al., 2018*; *Kim et al., 2019*). Furthermore, some studies have adopted strategies such as question ranking (*Zheng et al., 2018*), reinforcement learning (*Fan et al., 2018*), and multi-task learning (*Wang, Yuan & Trischler, 2017*) to further enhance the naturalness of the generated questions. And personalization is a comprehensive embodiment of other objectives. In summary, this study believes that diversification, controllability, and focus on teaching content should be the core objectives of QGEd.

In terms of literature breadth, this review covers more QGEd research articles than previous reviews, totaling 48. Regarding the review's scope, it is more comprehensive and systematic than previous work, covering the development of QG core technologies, the current state of research on QGEd's three core objectives, datasets, question evaluation schemes, and future research directions. In terms of perspective, this review starts from educational theories and practical needs, deeply discussing the achievements and shortcomings of current technical research, thereby providing better guidance for its future development. In summary, this review introduces existing QGEd technologies developed by other researchers, defines their core algorithms and distinctive features, and helps researchers gain a thorough understanding of the current state of QGEd development, thereby inspiring them to take the next steps in their work.

Intended audiences

This review is intended for two main groups of audiences with varying levels of expertise in computer. The first group consists of non-technical individuals, such as teachers and educational researchers, who may not have extensive experience with computer techniques. The second group includes technical audiences, such as education technology researchers and developers in the artificial intelligence industry, who may have more familiarity with these techniques. The first objective is to motivate teachers and educational researchers who have relied heavily on conventional question design and lack an understanding of computer vision algorithms. We aim to encourage them to engage with question design used in practical teaching and teaching research. The second objective is to assist and inform education technology researchers and developers about the current state of QGEd. This allows them to take inspiration from existing work and make modifications, thereby preventing them from having to reinvent the wheel.

The rest of this article is organized as follows. The methodology used to conduct this survey is discussed in 'Survey Methodology'. The progress of QG from the perspective of technological evolution and the evolution trends of QGEd are introduced in 'Development of Technology'. The research progress of QGEd in achieving the three core objectives is discussed in 'Three Core Objectives'. An overview of typical datasets commonly used for QG is presented from the perspective of QGEd in 'Datasets', followed by a review of commonly used evaluation schemes for QGEd in 'Evaluation Schemes'. Finally, discussions and future work directions are presented in 'Discussions', and the overall conclusions are summarized in 'Conclusion'.

CONTRIBUTIONS

This article makes several key contributions to the field of QGEd:

(1) We provide a systematic analysis of the research landscape of QGEd, tracing its evolution from heuristic rule-based methods to neural network technologies.

(2) We identify pivotal challenges and opportunities in QGEd, including the need for multimodal data processing, controllability of fine-grained cognitive and difficulty levels, and specialized educational dataset construction.

(3) We propose future research directions to address these challenges, emphasizing the development of automatic evaluation technologies and system architecture design.

(4) We offer a comprehensive overview of the field, synthesizing insights across its evolution, core objectives, datasets, and evaluation methods, which provides a clear understanding of the current state and progress for researchers and practitioners in educational technology.

SURVEY METHODOLOGY

This study was conducted following the guidelines of the 2018 PRISMA framework for scoping reviews (*Tricco et al., 2018*), which provides a set of rigorous and transparent methods to ensure trustworthy results. This review aims to collect recently published studies in order to analyze the development of QGEd research over the past decade (2015–2025). Due to the special timing of the manuscript, the studies published in 2025 have not been fully collected in this review.

Information source and search process

The five scientific databases were considered during the initial phase (see Fig. 1): Web of Science, IEEE Explore Digital Library, Springer Link, DBLP Computer Science Bibliography and ACM Digital Library. Each database includes relevant studies about QGEd. The search strategy included limiting the search results to studies that were published between 2015 and 2025 and that were written in English. The keywords chosen

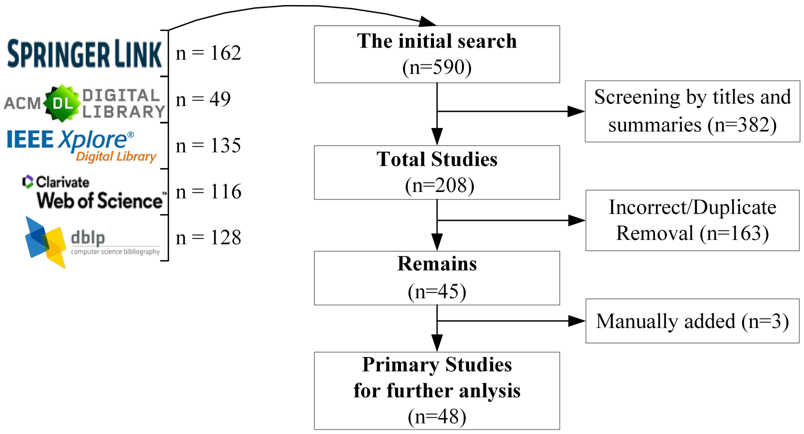

SPRINGER LINK
n = 162
ACM DL DIGITAL LIBRARY
n = 49
IEEE Xplore® Digital Library
n = 135
Clarivate Web of Science
n = 116
dblp computer science bibliography
n = 128
The initial search (n=590)
Screening by titles and summaries (n=382)
Total Studies (n=208)
Incorrect/Duplicate Removal (n=163)
Remains (n=45)
Manually added (n=3)
Primary Studies for further anlysis (n=48)

Figure 1 **Summary of scoping review databases.** Full-size DOI: 10.7717/peerj-cs.3203/fig-1

to identify the relevant records were "question generation," "question generator," "question generation for education," "difficulty-controllable generation," "question-type driven question generation," "diverse question generation," "Automatic education QG." It should be noted that this review primarily focuses on studies related to English QG, as the majority of the literature retrieved from citation databases centers on English. However, this does not imply that research on QG in other languages is excluded, which may also provide valuable insights and directions for future exploration.

Articles selection

The initial search on those five databases returned 590 articles. The titles and summaries of these articles were screened to exclude irrelevant works. As a result of this first screening, 208 articles were selected and considered for the next screening phase, which included a full review of the articles to check whether the eligibility criteria were met. Finally, 48 unique and fully accessible studies were identified as primary sources for further analysis in this review. Details about the number of records retrieved by each scientific database and the process of the data extraction and monitoring are shown in Fig. 1.

Data items

Each selected articles were indexed in a local database and, for each study, the following characteristics were included: title, year, core technology (heuristic rule, neural network or the combination of these two), core objective, database, evaluation metrics and evaluation/assessment strategies. Such characteristics were deemed relevant to reach the aim of the present work.

Synthesis of results

The authors of this article, based on the selected articles, provides a detailed and comprehensive review of the technological development trajectory, the current status of research on the three core objectives, commonly used datasets in research, and evaluation schemes of QGEd in the rest of sections. The basic information of the selected articles is

Table 3 The basic information of the surveyed papers about QGEd from January 2015 to May 2024.

Articles	Core technology	Diversification	Controllability	Focus on teaching content	Subject	Assessment scheme
Kumar, Banchs & D'Haro (2015)	Heuristic rule & NN	√			Biology	Manual
Olney, Pavlik & Maass (2017)	Heuristic rule	√	√	√	Medicine	Manual
Du, Shao & Cardie (2017)	NN		√		English	Manu+Auto
Yang et al. (2017)	NN	√	√		Multidisciplinary	Automatic
Du & Cardie (2017)	NN	√	√		English	Automatic
Song et al. (2018)	NN		√		English	Automatic
Flor & Riordan (2018)	Heuristic rule		√		English	Manual
Park, Cho & Lee (2018)	Heuristic rule	√			English	Manual
Gao et al. (2019)	NN		√		English	Manu+Auto
Du & Cardie (2018)	NN		√	√	English	Manu+Auto
Sun et al. (2018)	NN	√	√	√	English	Automatic
Wang et al. (2019)	NN		√		English	Manu+Auto
Chan & Fan (2019)	NN		√	√	English	Automatic
Wang et al. (2020a)	NN	√	√	√	English	Automatic
Liu et al. (2020)	NN	√			English	Manu+Auto
Steuer, Filighera & Rensing (2020)	NN			√	English	Manu+Auto
Jia et al. (2021)	NN		√	√	English	Manu+Auto
Srivastava & Goodman (2021)	NN		√		Language learning	Manu+Auto
Cheng et al. (2021)	NN		√	√	Multidisciplinary	Manu+Auto
Murakhovs'ka et al. (2021)	NN	√	√		Multidisciplinary	Manu+Auto
Cao & Wang (2021)	NN		√		Multidisciplinary	Manu+Auto
Qu, Jia & Wu (2021)	NN		√		English	Manu+Auto
Stasaski et al. (2021)	Heuristic Rule & NN		√		Multidisciplinary	Manu+Auto
Cui et al. (2021)	NN	√		√	Multidisciplinary	Manu+Auto
Steuer et al. (2021)	Heuristic Rule & NN			√	Multidisciplinary	Manual
Zou et al. (2022)	Heuristic Rule & NN	√	√		English	Manual
Ch & Saha (2023)	NN	√	√	√	Multidisciplinary	Manual
Zeng et al. (2022)	NN		√		Multidisciplinary	Automatic
Liu & Zhang (2022)	NN	√		√	Multidisciplinary	Manu+Auto
Yao et al. (2022)	Heuristic Rule & NN		√	√	Fairy Tale	Manu+Auto
Guo et al. (2023)	NN	√			Multidisciplinary	Manu+Auto
Gou et al. (2023)	NN	√			Multidisciplinary	Manu+Auto
Chomphooyod et al. (2023)	Heuristic Rule & NN	√	√		English	Manu+Auto
Matsumori et al. (2023)	NN	√	√	√	English	Manu+Auto
Prokudin, Sychev & Denisov (2023)	Heuristic Rule & NN		√		Programing	Manual
Bi et al. (2024)	NN	√	√		Multidisciplinary	Manu+Auto
Babakhani et al. (2024)	NN	√			Multidisciplinary	Manu+Auto
Bitew et al. (2024)	NN	√	√		Multidisciplinary	Manual

(Continued)

Table 3 (continued)

Articles	Core technology	Diversification	Controllability	Focus on teaching content	Subject	Assessment scheme
Duong-Trung, Wang & Kravcik (2024)	NN		√	√	Introduction to machine learning & neural networks and deep learning	Manu+Auto
Zhang, Xie & Qiu (2024)	NN		√	√	English	Manu+Auto
Moon et al. (2024)	NN		√		English	Manu+Auto
Teruyoshi, Tomikawa & Masaki (2024)	NN		√		English	Automatic
Tomikawa & Masaki (2024)	NN		√		English	Automatic
Tomikawa, Suzuki & Uto (2024)	NN		√		English	Manu+Auto
Tonga, Clement & Oudeyer (2024)	Prompt Engineering	√	√	√	Math	Automatic
Lee et al. (2024)	Prompt Engineering	√	√	√	English	Manu+Auto
Shoaib et al. (2025)	NN		√	√	Computer	Manu+Auto
Maity, Deroy & Sarkar (2025)	Prompt Engineering	√	√	√	Educational content	Manu+Auto

Note:
NN refers to Neural Network. Manu refers to Manual. Auto refers to Automatic.

also organized in the Table 3 for readers to quickly and conveniently locate the reference literature that meets their needs.

DEVELOPMENT OF TECHNOLOGY

The exploration of general-purpose AQG technologies has established a robust technical foundation for QGEd. This section is dedicated to categorizing these general QG techniques and summarizing their evolutionary trends, with the aim of offering future researchers a technical roadmap for QGEd studies.

Regardless of the type of QG task, from the perspective of technological evolution, they can generally be divided into three stages: heuristic rule-based methods, neural question generation (NQG) and prompt engineering.

Heuristic rule-based methods

Heuristic rule-based methods were the mainstream approach in early question generation. These methods typically rely on artificially designed transformation rules to convert given contexts into corresponding questions. They can generally be divided into three categories: template-based methods (*Mostow & Chen, 2009*; *Zheng et al., 2018*; *Fan et al., 2018*; *Liu et al., 2019*; *Fabbri et al., 2020*), grammar-based methods (*Wolfe, 1976*; *Kunichika et al., 2004*; *Mitkov & Ha, 2003*; *Heilman & Smith, 2010*; *Ali, Chali & Hasan, 2010*), and semantics-based methods (*Huang & He, 2016*; *Yao & Zhang, 2010*; *Flor & Riordan, 2018*; *Dhole & Manning, 2020*). Generally, these methods include two steps: content selection and question construction. First, the topic to be inquired about is selected through semantic or syntactic parsing methods, and then templates or transformation rules are used to convert the context of the selected topic into a natural language question.

Template-based methods

Template-based methods are commonly used for QG in closed-domain specific applications. For instance, *Zheng et al. (2018)* proposed a template technique for constructing questions from Chinese text. *Wolfe (1976)* designed a pattern matching method for use in educational systems. *Fabbri et al. (2020)* employed a template method to perform sentence extraction and question generation in an unsupervised manner.

Grammar-based methods

Grammar-based methods first determine the syntactic structure of the given text, and then apply syntactic transformation rules and the part-of-speech of interrogative sentences to generate questions. *Mitkov & Ha (2003)* were among the first to attempt generating multiple-choice questions using transformation rules. *Heilman & Smith (2010)* employed an overgeneration and ranking strategy for generating factual questions. *Ali, Chali & Hasan (2010)* used syntactic analysis, part-of-speech tagging, and named entity recognizers to generate questions.

Semantics-based methods

Semantics-based methods generate questions by performing semantic analysis on the text. *Huang & He (2016)* introduced lexical functional grammar (LFG) as a linguistic framework for QG. *Yao & Zhang (2010)* proposed a semantics-based method for QG based on minimal recursion semantic representation. *Lindberg et al. (2013)* used a template method incorporating semantic roles to generate questions that support online learning. *Flor & Riordan (2018)* also generated factual questions based on semantic roles. *Dhole & Manning (2020)* developed Syn-QG, which utilizes dependency relations, shallow semantic parsing, and custom rules to generate questions.

Neural question generation

In 2017, *Du & Cardie (2017)* first applied the sequence-to-sequence (Seq2Seq) encoder-decoder model to question generation. Since then, neural question generation (NQG) has become the mainstream approach in QG research, leading to a surge of high-quality studies and achieving more desirable results. Most of the work in NQG utilizes recurrent neural networks (RNN) or transformer networks with attention mechanisms to automatically learn patterns and relationships in the input text, capturing complex dependencies and generating questions that may not be explicitly defined in rules, adapting to different domains when there is sufficient training data available.

RNN-based methods

RNN-based models consist of an encoder and a decoder, both of which are RNNs, such as long short-term memory (LSTM). These models typically include a word embedding layer, an encoder layer, and a decoder layer, equipped with attention mechanisms and copying mechanisms. For instance, *Chali & Baghaee (2018)* incorporated the coverage mechanism into the RNN-based framework. *Tang et al. (2018)* introduced copying and post-processing mechanisms into the RNN-based framework to address the issue of rare or unknown words. Reddy et al. (2017) utilized the RNN-based framework to generate questions from a given set of keywords.

Transformer-based methods

The main difference between the Transformer and RNN-based models is that the transformer's encoder and decoder are entirely implemented by attention mechanisms. In answer-unaware QG, Kumar et al. (2019) proposed a cross-lingual model based on the Transformer encoder-decoder architecture to improve QG in the target language. Scialom, Piwowarski & Staiano (2019) suggested incorporating copying mechanisms, placeholder strategies, and Embeddings from Language Models (ELMo) into the Transformer. Pan et al. (2020) introduced contextual information and control signals into the Transformer-based Seq2Seq model to generate diversified Chinese questions from keywords. Using an answer as input for a QG system can generate more specific and context-related questions; such systems are known as answer-aware QG. In answer-aware QG, Chan & Fan (2019) employed the Transformer-based BERT model to address the QG task with answer position embedding. Chai & Wan (2020) proposed a Transformer-based approach that utilizes answer span information to generate questions in an autoregressive manner.

Pre-trained model based methods

As large-scale pre-trained language models have achieved remarkable results in various natural language processing tasks, fine-tuning pre-trained language models to adapt to downstream tasks has become a research hotspot. For instance, Dong et al. (2019) proposed a model called UniLM that is pre-trained using three language modeling tasks. Wang et al. (2020b) introduced a simple and effective knowledge distillation model called MiniLM for compressing large pre-trained models. Bao et al. (2020) utilized masking strategies to learn relationships between terms and pseudo-masking to learn relationships between segment contents. Qi et al. (2020) designed a pre-trained model aimed at simultaneously predicting n-gram terms, applying an n-stream attention mechanism. Xiao et al. (2020) proposed a multi-stage pre-training and fine-tuning framework called ERNIE-GEN. Wu et al. (2021) used a pre-trained BERT model for question type classification and integrated a penalty mechanism for QG. Yulan et al. (2022) addressed the "exposure bias" and "masking heterogeneity" issues in the decoding phase of pre-trained models by employing methods based on stochastic noise resistance and transfer learning, enhancing the adaptability of the pre-trained model UniLM in QG tasks.

Other neural-network based methods

In addition to the main QG methods mentioned above, some researchers have also employed machine learning methods such as graph models, generative adversarial networks, and variational autoencoders (VAEs). For instance, Liu et al. (2019) used graph convolutional networks to represent the syntactic dependency tree of paragraphs. Ma et al. (2020) constructed an answer-centric entity graph from documents and utilized graph convolutional networks (GCN) to aggregate semantic information at different

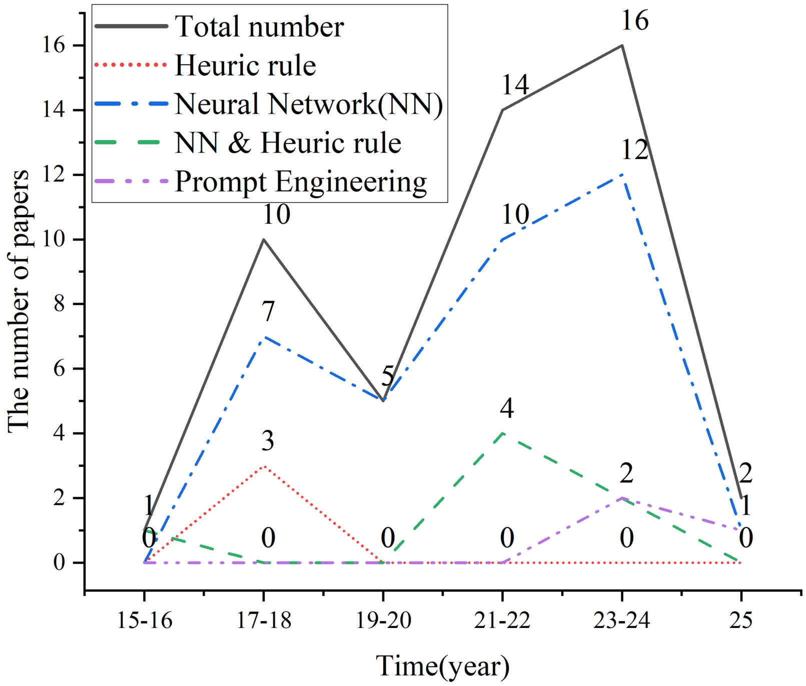


Figure 2 **The core technology evolution trends in QGEd.** Full-size DOI: 10.7717/peerj-cs.3203/fig-2

granularities. *Rao & Daumé (2019)* employed generative adversarial networks to refine QG. *Zhou et al. (2023)* introduced iterative graph neural networks (GNN) in the decoding stage to address the issue of losing rich sequential and semantic structural information from the paragraph during the decoding process.

Prompt engineering

The crux of prompt engineering lies in the meticulous design of input prompts to elicit outputs from LLMs that meet specific requirements. This approach obviates the need for updating the model's weights or parameters. Instead, it adeptly constructs prompts to steer the model towards generating questions in the desired direction (*Lee et al., 2024*; *Tonga, Clement & Oudeyer, 2024*; *Maity, Deroy & Sarkar, 2025*).

The core technology evolution trends in QGEd

To obtain high-quality data for the research, this study utilizes widely recognized scientific literature databases, including Web of Science, ACM, dblp, IEEE, and Springer. The search period is limited to January 2015 to February 2025, and a total of 48 QGEd research articles were collected. Based on the publication years of the literature, QGEd experienced a research peak between 2017 and 2018, as shown in Fig. 2. During this period, the number of studies using deep neural network models peaked at 7; in contrast, the number of studies based on heuristic rules rapidly decreased. After 2020, heuristic rules were primarily used to address the limitations of neural network models in controlling generation factors. Moreover, in recent years, prompt engineering has emerged as a novel and highly efficient method for guiding LLMs to automatically generate questions.

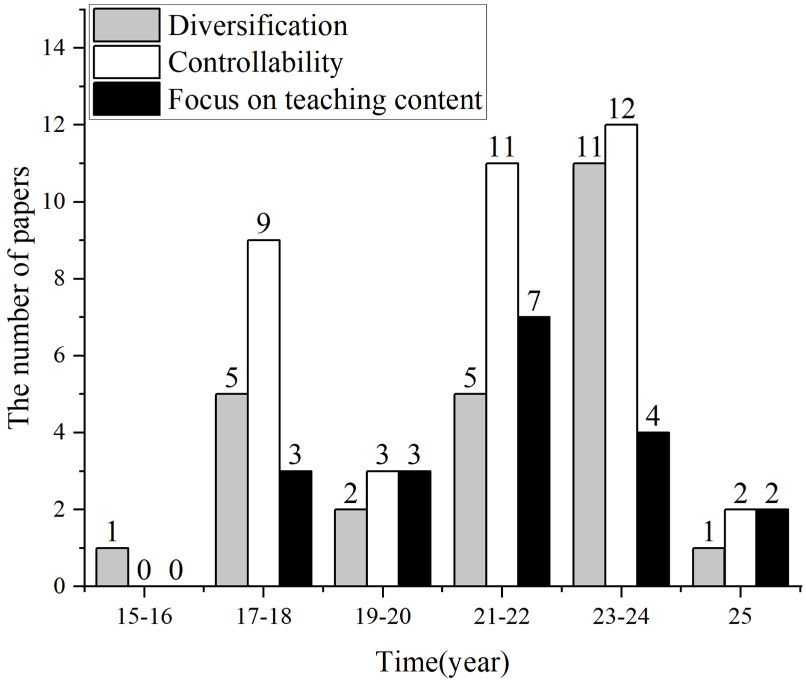


Figure 3 **Distribution of studies on the three core objectives.**

Full-size DOI: 10.7717/peerj-cs.3203/fig-3

THREE CORE OBJECTIVES

In this section, we will delve into the current state of development of QGEd with respect to its three core objectives: diversification, controllability, and focus on teaching content.

QGEd, as a subfield of QG specifically tailored for educational applications, has benefited from the technological advancements in QG, which have laid the groundwork for its implementation. However, the unique demands of the educational scenario have set higher standards for QGEd. As deep neural network models have become the main approach in this field, achieving the special goals of QGEd has gradually become possible. As shown in Fig. 3, the quality and density of research related to diversification, controllability, and focus on teaching content have significantly improved after 2017, with controllability receiving more attention from researchers than the other objectives.

Diversification

In the field of education, the diversification of question design is a widely recognized need. Diversified questions not only promote learners' thinking at different cognitive levels and deepen their understanding of the learning content, but also stimulate learning interest and creativity, and cultivate a multi-perspective approach to problem-solving. In the field of QGEd, question diversification is mainly reflected in two aspects: expression diversification and question type diversification (*Zhang & Zhu, 2021*). Expression diversification refers to conveying the same semantics through different expressions, and these questions, although stated in various ways, all point to the same answer. This kind of diversification helps to cultivate and exercise students' transferability and semantic discernment abilities. Question type diversification, on the other hand, involves the use of different types of questions, such as fill-in-the-blank, multiple-choice, short answer, and true/false questions, to meet different teaching needs and learners' learning styles.

Expression diversification

VAEs are widely adopted methods, representing a type of generative model for latent variables. Typically, related research utilizes an encoder to map observed variables (*i.e*., questions along with their corresponding answers and context) to latent variables, and then employs a decoder to map the latent variables back to observed variables. The training of this model is usually accomplished by maximizing the variational lower bound and some auxiliary objectives. For instance, *Wang et al. (2020a)* adopt a multimodal prior distribution strategy, selecting diversified background information related to the answer through a conditional variational autoencoder (CVAE), modeling it as continuous latent variables. They also explicitly model the expression types of questions, guiding the model to generate questions with diverse expressions through these auxiliary pieces of information. *Shinoda & Aizawa (2020)* proposed a variational QA pair generation model, introducing two independent latent random variables to model the one-to-many problem in QG. *Lee et al. (2020)* proposed a hierarchical conditional variational autoencoder (HCVAE), designing two independent latent variable spaces for questions and answers conditioned on input text.

Some studies simplify the QG task to generating diversified questions based on preset keywords to facilitate the generation process. These studies typically use keyword groups as input for the generation model, training it to complete the one-to-many QG task. For example, *Pan et al. (2020)* combine multi-head self-attention mechanisms with contextual information (questions that have already been generated, rather than the original text containing the keywords) and introduce control signals to increase question diversity while ensuring high relevance to the given keywords. *Liu et al. (2020)* adopt a similar approach, proposing an answer-clue-style aware question generation (ACS-QG) method. This method automatically constructs a dataset of (answer, clue, question word type) triplet structures from the SQuAD dataset to train the QG model. It further demonstrates the key role of different questioning clues (*i.e*., answer background information) in the diversity of question expression while enhancing the controllability of question expression methods.

Other studies have demonstrated that introducing external information or combining templates with neural network models can also benefit the generation of diversified questions. For instance, *Liu et al. (2017)* combined template-based methods with Seq2Seq learning to generate fluent and diverse questions. *Guo et al. (2023)* introduced a dual-model framework interwoven with two selection strategies, which extracts more diverse expressions from external natural questions and integrates them into the generation model to enhance question diversity. *Gou et al. (2023)* further considered the impact of external knowledge on question diversity, proposing a retrieval-enhanced type transformation framework (RAST) that utilizes type-diverse templates for question generation. *Zou et al. (2022)* proposed an unsupervised true/false QG method (TF-QG) that enables the QG model to create diverse judgment questions using heuristic templates generated by various NLP techniques.

Question type diversification

In the research on diversified question type generation, early heuristic rule-based question generation techniques deeply studied the grammar and semantics of sentences, enabling the generation of simple factual fill-in-the-blank questions and cloze tests from given texts (*Kumar, Banchs & D'Haro, 2015*; *Olney, Pavlik & Maass, 2017*). However, due to the difficulty of these strategies in overcoming the challenges of complex natural language processing (NLP), researchers began to seek alternative methods. With the significant advantages of deep neural networks in handling complex issues, more and more researchers are adopting neural network models to enhance the diversification of question types in QG. For example, *Park, Cho & Lee (2018)* used text embedding models to project documents, sentences, and word vectors into the same semantic space, thereby identifying sentences and words with educational value. After ranking the similarity of these elements, they were transformed into interrogative sentences and corresponding blanks for the interrogative words. This method can generate multiple-choice fill-in-the-blank questions without relying on any prior knowledge or additional datasets. *Chomphooyod et al. (2023)* used a text generation model trained with the Text-to-Text Transfer Transformer (T5) architecture to control the content and grammatical themes through keywords and part-of-speech (POS) templates. They also used rule-based algorithms to generate correct answers and distractors for 10 grammatical themes, thereby generating grammatical multiple-choice questions. The CLOZER algorithm introduced by *Matsumori et al. (2023)* utilizes a masked language model (MLM) and a new metric, gap score, to automatically generate educationally valuable open cloze questions (OCQ).

Controllability

Controllable QGEd primarily refers to the controllability of question difficulty and cognitive level. Difficulty and cognitive level are two interrelated but independent concepts. In the teaching process, it is necessary to design questions of different cognitive levels for the learning content based on teaching objectives, and different cognitive levels reflect different question difficulties. On the other hand, due to the varying actual ability levels of learners, the perceived difficulty of the same question may differ. Therefore, when generating questions, it is important to consider not only the cognitive level involved in the question but also the difficulty of the question.

Difficulty-controllable QGEd

The main principle of difficulty-controllable QGEd is to train the model on datasets with difficulty annotations for questions, using context information, answers, and difficulty as model inputs, and controlling the model to generate questions of a specified difficulty level by introducing relevant mechanisms. For example, *Shoaib et al. (2025)* manually annotated a large number of questions with difficulty levels and extensively collected questions that were already labeled with difficulty levels to construct a training dataset. It employed a conditional generative adversarial network (cGAN) model. By specifying conditions such as difficulty levels and subject domains during the generation process, and leveraging a balanced training dataset that covered multiple difficulty levels, the model was able to learn the characteristics of questions at different difficulty levels. This enabled the generation of multiple-choice questions (MCQs) that met predefined difficulty requirements.

However, the cost of manually annotating a large number of questions with difficulty levels is substantial. As a result, more research has been inclined towards designing an automatic annotation scheme for question difficulty. *Gao et al. (2019)* utilized two distinct machine reading comprehension (MRC) models to simulate test-takers, labeling the difficulty of questions based on whether both MRC models could correctly answer the same question. This approach led to the construction of the first difficulty-annotated dataset based on the SQuAD dataset, and it enabled the implementation of difficulty-controllable QG. Although the difficulty annotation of questions in this study is straightforward, it significantly reduces costs and is relatively objective, providing a new perspective for constructing difficulty-annotated datasets.

For example, *Zhang, Xie & Qiu (2024)* trained various pre-trained language models (PLM) models on the CLOTH dataset to simulate test-takers of different proficiency levels, and the effectiveness of the difficulty control strategies is evaluated using IRT models. They presented a framework for generating difficulty-controllable multiple-choice cloze tests (MC cloze test) using PLM and item response theory (IRT). The framework estimates gap complexity *via* the Shannon entropy of candidate gap words and selects distractors of varying difficulty levels using PLM confidence score, semantic similarity, and Levenshtein distance. It also incorporates rules to reduce the generation of invalid distractors.

In other related research, multiple question-answering systems were trained to act as proxies for real learners. These systems were based on a variety of pre-trained models, such as BERT, RoBERTa, and DeBERTa, and were fine-tuned using data subsets of different magnitudes. The performance of these question-answering systems on the dataset was leveraged to determine the difficulty of the questions *via* item response theory (IRT). This calculated difficulty was a key factor in training a question-generation model that can produce questions with controllable difficulty (*Moon et al., 2024*; *Teruyoshi, Tomikawa & Masaki, 2024*; *Tomikawa & Masaki, 2024*; *Tomikawa, Suzuki & Uto, 2024*).

Some research has also attempted to incorporate knowledge tracing techniques, using the historical test results and reading comprehension question texts of specific learners to automatically predict and control the difficulty of generated questions (*Srivastava & Goodman, 2021*). These approaches must rely on specific subjects (automated models or real test-takers), and their results can vary greatly due to the different characteristics of the subjects, leading to limited scalability and hindering the exchange and comparison between different studies.

The absolute difficulty of a question is primarily measured by the characteristics of the question itself, without being influenced by the subject taking the test. In such studies, *Cheng et al. (2021)* defined the difficulty of a question as the number of reasoning steps required to answer it. They used graph-based methods to model logic, first generating an initial simple question based on the first node of the reasoning chain with a question generator (QGInitial), and then gradually increasing the reasoning steps through a question rewriting module (QGRewrite) to transform the simple question into a more complex one. On this basis, they further increased the difficulty of the question by using two rewriting patterns: Bridge and Intersection. *Bi et al. (2024)* proposed a simple and intuitive cross-domain automatic difficulty evaluation technique that focuses solely on the complexity of the question itself, significantly reducing the impact of subjective factors from test-takers. This approach is conducive to the unification of question difficulty evaluation standards and greatly enhances the generalizability of absolute difficulty evaluation. Research on multiple-choice question generation, on the other hand, focuses on increasing the similarity of distractors to enhance the absolute difficulty of the questions (*Ch & Saha, 2023*; *Murakhovs'ka et al., 2021*).

Cognitive level controllable QGEd

Cognitive level controllability emphasizes that questions should be able to cover different cognitive levels, from basic memory and understanding to higher levels of analysis, evaluation, and creation. When generating questions, it is important to consider the cognitive development stage of students and design questions that are appropriate for that stage. For example, for beginners, more foundational questions can be provided to consolidate knowledge; for advanced learners, more complex questions that require deep thinking can be designed to promote the development of their critical thinking and innovative abilities. There are two main schools of thought in academia regarding the classification of cognitive levels of questions. One is the delineation method proposed by *Graesser & Person (1994)*, and the other is the well-known Bloom's taxonomy of questions (*Bloom, 1956*). The former, based on extensive research of questions in psychology, education, and artificial intelligence, and according to the type of expected answers for educational purposes, identified sixteen common types. The latter classifies questions based on the cognitive processes required to obtain answers.

In both Graesser's and Bloom's classification methods, each question type is further divided into three common difficulty levels: low, medium, and high, as detailed in Table 4 (*Al Faraby, Adiwijaya & Romadhony, 2024*). For ease of understanding and discussion, these questions can further be generalized into factual and non-factual questions. Factual questions aim to retrieve specific factual information, such as names, locations, or time-related details. These types of questions are usually directly converted from specific sentences or information, and their answers can be found directly from the given context, involving relatively simple cognitive activities. Non-factual questions often require long and abstract answers, such as open-ended questions, detail reasoning questions, comparative analysis questions, *etc*. This category of questions tends to involve higher-order cognitive activities. The answers to non-factual questions may be in the given context or may not be in the given context.

1. *Factual QGEd*

Factual questions are the most commonly generated in QG research, and the technology is relatively mature. Early heuristic rule-based methods have been able to generate smooth, relevant, accurate, and controllable questions to a certain extent, with strong

Table 4 The correspondence between cognitive level and difficulty levels.

Difficulty	Graesser	Bloom
LOW	Verification Disjunctive Concept completion Example	Recognition Recall
MIDDLE	Feature specification Quantification Definition Comparison	Comprehension
HIGH	Interpretation Causal antecedent Casual consequence Goal orientation Instrumental/Procedural Enablement Expectation Judgmental	Application Analysis Synthesis Evaluation

interpretability. For example, *Wolfe (1976)* designed a pattern matching method for educational systems. *Lindberg et al. (2013)* incorporated semantic roles into template methods to generate questions supporting online learning. *Flor & Riordan (2018)* also generated factual questions based on semantic roles. However, these methods rely on domain expertise and experience, and their effectiveness depends on the variety and quantity of predefined rules or templates, making it difficult to handle more diverse and complex contexts that were not anticipated during the rule design process. Due to the inherent complexity of language, it is almost impossible to manually discover and summarize all question generation rules, which would require a huge manual effort. And rules designed for one domain are usually difficult to quickly transfer to other domains. Additionally, the construction of manual templates will also greatly limit the diversity of generated questions. Therefore, these methods are difficult to generalize and apply on a large scale.

With the advent of the deep learning era, neural network models quickly replaced heuristic rule-based methods. *Du & Cardie (2017)* were among the first to introduce attention mechanism-based sequential learning models into QG research, directly enabling models to generate questions based on the sentences containing the answers through Seq2Seq learning, demonstrating that NQG techniques outperform traditional heuristic rule-based approaches. However, the aforementioned appoach has two major issues: (1) It neglects the processing of specific answer information within sentences. (2) It neglects the processing of contextual information surrounding the sentences. These issues prevent the scheme from ensuring the relevance of generated questions to answers and context, leading to poor interpretability, and an inability to handle paragraph-level QG tasks.

To address issue (1), researchers have chosen to embed the answer location information into neural network models. For example, *Zhou et al. (2017)* encoded the BIO tags of the answer locations as real-valued vectors embedded into the encoder layer. *Zhao et al. (2018)* proposed a gated self-attention network encoder with answer marking and a maximum output pointer to avoid generating repetitive questions. Yang et al. (2017) and Yuan et al. (2017) respectively added an additional binary feature to the word embeddings of paragraph markers and document markers. Steuer, Filighera & Rensing (2020) used answer-aware NQG technology, combining two new answer candidate selection strategies based on semantic graph matching, to generate clear, answerable, and educationally valuable factual questions. Although these approaches have made the relevance of generated questions to answers more intimate, they struggle to ensure the relevance of questions to context, and there is a problem of generating questions that contain the answer itself.

To address the aforementioned issues and issue (2), researchers have chosen to independently encode the answer and context and model their relationship, enabling the model to generate high-quality factual questions relevant to both answers and context. For instance, Kim et al. (2019) used a dual encoder to separately encode the answer and the paragraph text with answer masking, and during the question generation process, they used an Answer-Separated Decoder to combine context features from the paragraph (obtained through attention mechanisms) and key features from the target answer's keyword network, solving the problem of generating questions that contain answer information and improving the relevance of questions to paragraph context. Wang et al. (2019) designed a weakly supervised discriminator that encodes the answer and context separately to capture the relationship between the two. Song et al. (2018) used two independent bidirectional long short-term memories (Bi-LSTMs) to encode the context and answer, and then matched each hidden state in the paragraph with all hidden states of the answer through a multi-perspective context matching (MPCM) algorithm to gather context information related to the answer. Zeng et al. (2022) proposed a dual-attention-based paragraph-level QG model, which, although only encoding the overall paragraph, used attention mechanisms to capture the relationship information between the answer and the overall paragraph for both the paragraph and the sentence containing the answer.

2. *Non-factual QGEd*

The main approach in this type of research currently is to reconstruct or modify datasets to optimize input features for training specific models to generate non-factual questions. For example, inspired by Graesser's classification, Cao & Wang (2021) directly adopted the cognitive-based question classification scheme revised by Olney, Graesser & Person (2012) to construct a new dataset and successfully implemented technology for generating different cognitive levels of questions; Jia et al. (2021) reconstructed the original RACE dataset and achieved certain results in generating English examination questions; Murakhovs'ka et al. (2021) built a new QA dataset by mixing nine datasets with different answer types to train the MixQA model, successfully generating a diversified cognitive levels of questions; Qu, Jia & Wu (2021) compared and analyzed the RACE and SQuAD datasets in terms of question form structure, concluding that the RACE dataset is more suitable for QGEd, and built a model specifically for generating non-factual questions trained on the RACE dataset; Stasaski et al. (2021) used a causality extractor to construct a causal question dataset to adjust neural networks, successfully generating causal questions from textbook texts.

There are also studies attempting to generate non-factual questions by modifying neural network models, such as (*Babakhani et al., 2024*) who fine-tuned the architectures of large language models flan-T5 and GPT-3, using Seq2Seq generation technology to automatically generate subjective questions for news media posts. *Qu, Jia & Wu (2021)* proposed a multi-agent communication model to iteratively generate and optimize questions and keywords, then use the generated questions and keywords to guide the generation of answers. However, the schemes for modifying neural network models to generate non-factual questions still do not have controllable specific cognitive levels. Therefore, datasets with cognitive level labels are of great significance for controllable QGEd research.

Some studies have leveraged the powerful content learning capabilities of LLMs, employing fine-tuning or prompt engineering to train or guide these models in generating high-quality questions. *Duong-Trung, Wang & Kravcik (2024)* fine-tuned ChatGPT-3.5-turbo using a training dataset comprising 1,026 master's course questions across 29 subjects, resulting in the development of BloomLLM. These data were annotated by experts according to the six cognitive levels of Bloom's Taxonomy, ensuring that the generated questions are semantically interrelated and aligned with educational objectives. Through this approach, BloomLLM is capable of generating questions that better meet educational needs, addressing the shortcomings of basic LLMs in semantic relevance and hallucination issues, thereby enabling more effective application of generative AI in the field of education. In terms of prompt engineering, *Lee et al. (2024)* developed a framework based on a 2D matrix, in conjunction with prompt engineering techniques for ChatGPT, to facilitate the automatic generation of English reading comprehension questions. Teachers can select appropriate question types and formats from the matrix according to the reading materials and learning objectives, and input corresponding prompts. ChatGPT will then generate the relevant questions. The generated questions cover a range of types, including literal comprehension and inferential judgment. *Tonga, Clement & Oudeyer (2024)* leveraged LLMs to simulate the roles of students and teachers. By designing specific prompts, it guided the teacher model to generate probing questions. The diversity and creativity of the generated prompts were controlled by adjusting the model's temperature parameter. These questions were heuristic in nature.

Focus on teaching content

A question may belong to a higher cognitive level, but this does not automatically endow it with educational value. Questions with educational value should enable learners to gain knowledge, develop skills, and achieve teaching objectives or fulfill the purposes of teaching evaluation in the process of answering the questions. To generate questions with educational value, and to make the generated questions more educationally valuable, some studies focus on generating questions targeted at content with educational value within the context, and use richer information related to the answers as input to improve the quality of question generation, such as the external knowledge background of the answers, semantic features, and positional features.

Regarding the introduction of external knowledge background information, *Liu & Zhang (2022)* utilized the Chinese Wikipedia knowledge graph (CN-DBpedia2) to extract knowledge triples related to the answers and used the pre-trained language model BERT to encode the target answers and the graph separately, enriching the answer feature information and enhancing the educational value of the generated questions. In terms of utilizing the semantic features of answers, *Zhou et al. (2017)* incorporated linguistic features such as part-of-speech, named entities, and word forms into the QG system's input; *Du & Cardie (2018)* improved the accuracy of questions by annotating each pronoun's antecedent and its coreference location features in the input text. Regarding the processing of answer location features, *Sun et al. (2018)* designed a position-aware model that enhances the model's perception of the context word positions by encoding the relative distance between context words and answers as position embeddings; *Cheng et al. (2021)* utilized open information technology to construct the original text into a directed graph composed of entity nodes and relationship edges, and then encoded it with graph neural networks, thereby leveraging the positional information of the answer nodes within the graph.

However, the above technologies, which focus on teaching content, all adopt answer-aware QG strategies, requiring the target answers to be manually annotated in advance. This undoubtedly increases the workload of human teachers in practical applications, reflecting the issue of insufficient technological intelligence level. Therefore, technologies that can generate answers and corresponding questions based on a given paragraph have been valued by researchers. This technology is known as joint question-answer generation (*Cui et al., 2021*). In joint question-answer generation, automatically selecting content with educational value from a given context is a key step. For example, *Chen, Yang & Gasevic (2019)* proposed nine strategies for selecting important sentences, including choosing sentences at the beginning of a paragraph, sentences containing new words, and sentences with the lowest similarity to other sentences. *Steuer et al. (2021)* used keyword filtering and classifiers to filter out sentences containing key concepts, and then used dependency parsing and semantic graph matching techniques to identify and extract content with complete semantic information and educational value from textbooks as target answers for generating questions. *Yao et al. (2022)* used heuristic rules to select key words that help understand the story, such as characters, emotions, and causal relationships. *Ch & Saha (2023)* identified key content in textbooks through word frequency statistics. In addition to rule-based and statistical methods, there are also studies that train models to automatically select sentences. For example, *Du & Cardie (2017)* used a hierarchical neural network framework to encode the input context at both the sentence level and paragraph level, treating sentences containing the target answer as the target sentences, and the given context as input, directly training a neural network model capable of automatically extracting sentences with question-asking value.

Recently, *Maity, Deroy & Sarkar (2025)* directly utilized education-related content as the data source, integrating retrieval-augmented generation (RAG) based on BART with prompt engineering for large language models. By combining the strengths of retrieval and few-shot learning, it generated higher-quality questions that better aligned with educational objectives.

DATASETS

In this section, we will present an overview of the datasets frequently employed in QG and provide detailed introductions and summaries of several representative datasets.

The availability and characteristics of datasets are crucial for NQG systems. Based on the source of the dataset content, this study categorizes the surveyed public datasets into two major categories: non-education-related and education-related. Table 5 shows a comparison of these datasets in terms of the main cognitive levels of questions, question types, subjects (content sources), the presence of cognitive/difficulty classification labels, and the forms of given contexts. Specifically, in the cognitive level column, datasets labeled as containing non-factual questions indicate that the datasets include such questions, but this does not imply that all questions within the datasets are non-factual.

Non-education related datasets

The content of non-education-related datasets is primarily sourced from Wikipedia, news articles, speech transcripts, or paragraph texts from the internet, rather than educational textbooks. Even so, some datasets that encompass non-factual questions and difficulty classification features also offer valuable resources for exploring how to generate questions that are of higher cognitive levels or differentiated in difficulty. For instance, HotpotQA (*Yang et al., 2018*) includes two modes: simple and difficult, with solving difficult mode questions requiring cross-document reasoning. NQ (*Kwiatkowski et al., 2019*) is divided into “simple, medium, difficult” difficulty levels based on the naturalness of the questions and the findability of the answers. ODSQA (*Lee et al., 2018*) may require reasoning and understanding based on colloquial questions. The questions in CoQA (*Reddy, Chen & Manning, 2019*) are generated based on a segment of situational dialogue stories, forming a logically coherent sequence of questions. This type of question is particularly beneficial for cultivating learners’ language logical expression abilities. The questions in Cosmos QA (*Huang et al., 2019*) require common sense reading comprehension, mainly including questions about the possible causes or consequences of events.

Education related datasets

Education related datasets not only focus on content that can be used for education as their data source but also more universally pay attention to the collection and processing of questions with different difficulties and more cognitive levels. Among them, ARC (*Clark et al., 2018*), although it includes a variety of question types, only simply divides the questions into an Easy set and a Challenge set. RACE (*Lai et al., 2017*), as described in section “THREE CORE OBJECTIVES,” has achieved considerable results in various QGEd studies and is divided into two difficulty levels: the middle school RACE-M and the high

Table 5 **The datasets for QG including QGEd.**

Datasets		Cognitive levels	Question types	Subjects (content sources)	Cognitive/difficulty classification	Language
Non-education related	SQuAD (*Rajpurkar et al., 2016*)	Factual	Short answer	Wikipedia	×	English
	HotpotQA (*Yang et al., 2018*)	Non-factual	Short answer	Wikipedia	√	English
	NewsQA (*Trischler et al., 2017*)	Factual	Short answer	Wikipedia	×	English
	MLQA (*Lewis et al., 2020*)	Factual	Short answer	Wikipedia	×	Multilingual
	XQuAD (*Artetxe, Ruder & Yogatama, 2020*)	Factual	Short answer	Wikipedia	×	Multilingual
	ODSQA (*Lee et al., 2018*)	Non-factual	Short answer	Speeches	×	English
	NQ (*Kwiatkowski et al., 2019*)	Non-factual	Short answer	Wikipedia	√	English
	Cosmos QA (*Huang et al., 2019*)	Non-factual	Multiple choice	Common sense	×	English
	CMRC (*Cui et al., 2019*)	Factual	Short answer	Wikipedia	×	Chinese
	ELI5 (*Fan et al., 2019*)	Non-factual	Short answer	Web document	×	English
	CoQA (*Reddy, Chen & Manning, 2019*)	Non-factual	Short answer	Conversations	×	English
	DuReader (*He et al., 2018*)	Non-factual	Short answer	Web document	×	Chinese
Education related	OpenBookQA (*Mihaylov et al., 2018*)	Non-factual	Multiple choice	Science	×	English
	SciQ (*Welbl, Liu & Gardner, 2017*)	Factual	Multiple choice	Science	×	English
	MedMCQA (*Pal, Umapathi & Sankarasubbu, 2022*)	Non-factual	Multiple choice	Medicine	×	English
	RACE (*Lai et al., 2017*)	Non-factual	Multiple choice	Entrance exam	√	English
	RACE_C (*Liang, Li & Yin, 2019*)	Non-factual	Multiple choice	English exam	√	English
	EQG_RACE (*Jia et al., 2021*)	Non-factual	Multiple choice	English exam	√	English
	EMBRACE (*Zyrianova, Kalpakchi & Boye, 2023*)	Non-factual	Multiple choice	English exam	√	English
Education related	ReClor (*Yu et al., 2020*)	Non-factual	Multiple choice	English exam	×	English
	TabMCQ (*Jauhar, Turney & Hovy, 2016*)	Non-factual	Multiple choice	Multidiscipline	×	English
	EduQG (*Hadifar et al., 2023*)	Non-factual	Multiple choice	Multidiscipline	×	English
	CLOTH (*Xie et al., 2018*)	Non-factual	Fill-in-the-blank	Middle school test	√	English
	INQUISITIVE (*Ko et al., 2020*)	Non-factual	Short answer	Multidiscipline	×	English
	ARC (*Clark et al., 2018*)	Non-factual	Multiple choice	Science	√	English
	MAWPS (*Koncel-Kedziorski et al., 2016*)	Non-factual	Math word	Math	×	English
	Math23K (*Wang, Liu & Shi, 2017*)	Non-factual	Math word	Math	×	Chinese
	CSFQGD (*Zhang et al., 2021*)	Factual	Fill-in-the-blank	Chinese test	×	Chinese
	SocratiQ (*Ang, Gollapalli & Ng, 2023*)	Non-factual	Short answer	Multidiscipline	×	English
	LearningQ (*Chen et al., 2018*)	Non-factual	Short answer	Multidiscipline	√	English
	FairytaleQA (*Xu et al., 2022*)	Non-factual	Short answer	Fairy tale	√	English

school RACE-H. CLOTH (*Xie et al., 2018*) categorizes questions into four types based on the cognitive effort required to answer them: Grammar, Short-term reasoning, Matching/ Paraphrasing, and Long-term reasoning, with the largest proportion of Long-term reasoning questions. FairytaleQA divides questions into explicit and implicit categories based on whether the answers can be directly found in the text. In fact, the former belongs to factual questions, whereas implicit questions are non-factual questions that require summarization and judgment based on implicit information in the text.

Due to the large scale and high-quality real exam questions of RACE, it subsequently became one of the most commonly used datasets in QGEd, and some follow-up work has further supplemented RACE. In 2019, *Liang, Li & Yin (2019)* released the RACE-C, in which they collected real exam questions from Chinese university English reading comprehension, further enriching the difficulty range of RACE's questions. In 2021, based on RACE, *Jia et al. (2021)* released EQG-RACE specifically for educational QG tasks. EQG-RACE removed general questions not suitable for QG tasks and automatically annotated the key sentences of the questions by calculating the ROUGE-L between the answers and the article sentences. In 2023, *Zyrianova, Kalpakchi & Boye (2023)* conducted a detailed evaluation of RACE by scoring some data samples and manually annotated the key sentences of some questions, releasing EMBRACE.

Although these datasets have potential for QGEd research, the division of cognitive levels and difficulty in most datasets is empirical, lacking unified, clear, and quantifiable standards, which is not conducive to the development of controllable QGEd with cognitive levels and difficulty. To achieve more fine-grained controllable QGEd, a more specific and scientific classification and annotation of the cognitive levels and difficulty of questions are still needed. For example, LearningQ (*Chen et al., 2018*) uses the Bloom's Taxonomy introduced in section "THREE CORE OBJECTIVES" to annotate and categorize questions, but only 200 random questions in the dataset have this classification label, which is far from meeting the training needs of NQG.

In fact, according to section "THREE CORE OBJECTIVES," recent outstanding research in the field of QGEd has conducted extensive work on self-built datasets. These efforts essentially involve either automatic or manual annotation of training data with difficulty levels or cognitive hierarchies. This highlights the significant role of question difficulty and cognitive hierarchy, as two educational attributes, in the construction of datasets specifically dedicated to QGEd research.

EVALUATION SCHEMES

In this section, we focus on presenting the main evaluation schemes and criteria for QGEd systems. Our article primarily introduces these criteria rather than analyzing the State-of-the-Art (SoTA) outcomes, which are the results obtained through these criteria. We believe that establishing clear evaluation standards is crucial as it provides a unified framework for measuring progress and innovation. This allows researchers to better understand and compare different approaches within the field. There are two commonly used evaluation schemes in QGEd from two aspects: automatic schemes and manual schemes.

Question quality evaluation is a critical component in determining the success or failure of QGEd research. This section provides a brief overview of the current automatic and manual schemes used for evaluating QGEd systems.

Automatic schemes

Currently, most QGEd studies employ automatic evaluation metrics inherited from traditional QG research, with BLEU (*Papineni et al., 2002*), ROUGE (*Lin, 2004*), METEOR (*Banerjee & Lavie, 2005*), chrF (*Popović, 2015*) and BERTScore (*Zhang et al., 2020*) being the primary ones.

1. *BLEU*

BLEU is an n-gram precision-based evaluation metric that evaluates quality of generated text by comparing the n-gram overlap between generated text and a set of reference texts, with the specific calculation method as shown:

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{n} w_n \ \log \ p_n\right). \tag{2}$$

Here, p_n represents the precision of the nth gram, w_n is the corresponding weight, which is typically a uniform weight in BLEU, meaning that the weight of each n-gram is the same. BP stands for the brevity penalty factor, which comes into effect when the length of the generated text is less than or equal to the length of the reference text.

2. *ROUGE*

ROUGE evaluates the quality of generated text by comparing the overlapping units (such as n-grams, word sequences, and word pairs) between machine-generated text and ideal reference texts. It includes a total of four metrics: ROUGE-N, ROUGE-L, ROUGE-W, and ROUGE-S. The value of ROUGE-N is obtained by dividing the number of shared n-grams by the total number of n-grams in the reference summary. This ratio reflects the similarity in content between the generated text and the reference text. The specific calculation method is shown as:

$$\mathrm{ROUGE} - \mathrm{N} = \frac{\sum_{S\in\{RefrenceSummaries\}} \sum_{gram_n} Count_{match}(gram_n)}{\sum_{S\in\{RefrenceSummaries\}} \sum_{gram_n} Count(gram_n)}. \tag{3}$$

This formula calculates the n-gram recall rate, which is the ratio of the number of common n-grams between the machine-generated text and the reference text to the total number of n-grams in the reference text.

ROUGE-L is an evaluation method based on the longest common subsequence (LCS). It seeks the longest sequence of identical words between two sequences (generated text and reference text). The longer the LCS, the more similar the two texts are. ROUGE-L uses the F-measure (a metric that combines precision and recall) to evaluate similarity. ROUGE-W is an evaluation method for weighted longest common subsequence (weighted LCS). It is similar to ROUGE-L but adds weighting for consecutive matching words, meaning that if two words are consecutively matched, they will receive a higher score. ROUGE-S is based on the statistics of Skip-Bigrams. Skip-Bigrams are pairs of words that appear in a sentence but can have any number of other words in between. In summary, ROUGE focuses more on the recall of correct information in the generated text compared to BLEU when evaluating the similarity between generated text and reference text, that is, how much information from the reference text is included in the generated text.

3. *METEOR*

METEOR is an evaluation metric that takes into account synonyms, morphological variations, and sentence structure. Its calculation is more complex, involving the alignment of words in the generated text and the reference text, and calculating a score based on matches. The calculation formula is shown as:

$$\begin{aligned}&Fmean = \tfrac{10PR}{R+9P}\\&Penalty = 0.5 * \left(\frac{\#chunks}{\#unigrams_{matched}}\right)^3\\&\text{METEOR Score} = Fmean * (1 - Penalty).\end{aligned} \tag{4}$$

METEOR can calculate the similarity between generated questions based on word-level precision and recall, as well as a penalty for word order. The *Fmean* is the harmonic mean of unigram precision *P* and unigram recall *R*, with a higher weight given to recall as *9R*. The *Penalty* calculates the degree of fragmentation of matches due to word order disruption; if words in the generated text match with words in the reference text but are out of order, the *Penalty* increases.

4. *ChrF*

ChrF (Character n-gram F-score) is an evaluation metric based on character-level assessments, primarily used to evaluate the quality of text generation tasks. It assesses the similarity between reference and generated texts by calculating the matching counts of character-level n-grams. The core advantages of chrF include: (1) It can better capture syntactic and semantic information in morphologically rich languages; (2) It is language-independent, as it does not rely on language-specific tokenization processes and is applicable to multiple languages; (3) It is both simple and efficient, requiring no additional tools or knowledge sources. The general formula for the chrF score is:

$$\text{chrF}\beta = \left(1 + \beta^2\right)\frac{\text{chrP} \cdot \text{chrR}}{\beta^2 \cdot \text{chrP} + \text{chrR}}. \tag{5}$$

Here, chrP and chrR represent the precision and recall of character n-grams, calculated as the average across all character n-grams. The parameter β is used to adjust the relative importance of recall compared to precision, specifically giving β times more weight to recall. This metric is language-independent and token-agnostic, making it highly versatile and widely applicable across various translation tasks.

5. *BERTScore*

BERTScore is a text similarity evaluation metric based on the BERT pre-trained model, designed to measure the semantic similarity between generated text and reference text. It leverages contextual embeddings, utilizing the context-aware word vectors generated by the BERT model to capture semantic information. This approach enables BERTScore to demonstrate strong semantic matching capabilities, outperforming traditional evaluation methods based on n-gram overlap (such as BLEU) in capturing semantic similarity. Additionally, BERTScore supports evaluation across multiple languages, further enhancing its applicability in diverse linguistic contexts.

The calculation process of BERTScore is as follows: (1) Encode the candidate and reference texts using the BERT model to obtain contextual embeddings for each word; (2) Calculate the cosine similarity between each word in the candidate text and each word in the reference text; (3) Use a greedy matching algorithm to find the most similar word in the reference text for each word in the candidate text; (4) calculate precision, recall, and F1-score based on the matching results. Precision quantifies the proportion of words in the generated text that match the reference text, while recall measures the proportion of words in the reference text that match the generated text. The F1-score serves as a balanced metric, taking into account both precision and recall.

6. *Other metrics*

The aforementioned several metrics are commonly used in traditional QG research to evaluate the natural fluency of generated questions, focusing on the relevance and answerability of the generated questions, but they cannot directly evaluate the cognitive levels, difficulty, and diversification of the questions, which are the quality dimensions of more concern in the field of education. Researchers, considering this, have begun to explore other metrics for automatically evaluating the quality of generated questions. In terms of evaluating the cognitive level of questions, *Stasaski et al. (2021)* designed the Cause/Effect Presence metric to specifically evaluate the quality of causal questions. When the expected answer to a generated question is a "cause," the Cause Presence metric checks whether the "effect" extracted from the text is mentioned in the question. For questions where the "cause" is the expected answer, the size of the intersection between the words of the extracted "effect" appearing in the question and the set of "cause" words is calculated, then divided by the size of the "cause" word set to obtain the cause recall, and the same goes for Effect Presence. *Steuer, Filighera & Rensing (2020)* used the frequency of Wh-words to make a rough evaluation of the cognitive level of questions.

For automatic difficulty evaluation, please refer to section "Difficulty-controllable QGEd," which can be divided into absolute difficulty evaluation (*Gao et al., 2019*; *Srivastava & Goodman, 2021*) that only considers the complexity of the question itself, and relative difficulty evaluation (*Bi et al., 2024*; *Cheng et al., 2021*) that is calculated in combination with the test-taker's ability.

It is mainly evaluated by inversely evaluating similarity to evaluate the diversification of the questions, that is, the degree of difference in expression. For example, if a generated question is not similar to any question in the reference set, it can be considered novel. *Srivastava & Goodman (2021)* used novelty to evaluate question quality, which essentially uses text similarity metrics (such as cosine similarity, Jaccard similarity, *etc*.) to evaluate the difference between each generated question and the questions in the reference set. *Wang et al. (2020a)* and *Gou et al. (2023)* both used the pairwise metric to calculate the average BLEU-4 score between any two questions in the generated question set. A lower pairwise indicates a higher degree of question diversity.

In summary, automatic schemes focus on evaluating the similarity between the generated questions and the target questions, while relying on manual evaluation for quality dimensions more relevant to education, such as difficulty, cognitive level, and diversity. This inevitably introduces subjectivity from the evaluator, affecting the objectivity of the evaluation, which is not conducive to comparison and exchange among similar studies, and also increases research costs and time, affecting research efficiency. Although some studies have designed other automatic evaluation schemes for the cognitive level and diversity of questions, these schemes are still based on semantic similarity evaluation, whereas factors such as the complexity of knowledge, the degree of association between question answers and context have a greater impact on cognitive levels than semantic similarity between questions.

Manual schemes

Automatic evaluation holds significant advantages in terms of research efficiency and cost. However, existing reliable evaluation strategies are largely confined to the realm of semantic similarity assessment. When it comes to evaluating the quality of questions for educational applications, greater emphasis should be placed on dimensions such as educational value embedded in the questions. In this regard, human evaluation by educational experts still plays an irreplaceable role.

Manual evaluation typically covers three main criteria: naturalness (fluency, relevance, and answerability), difficulty, and helpfulness (educational value), which may be expressed differently or include other standards in different studies. The evaluation method usually involves experts, crowd-sourced participants, or the authors themselves, who annotate the generated questions based on specific metrics, and finally, the evaluation results are derived through statistics. For example, in the research (*Gao et al., 2019*), the difficulty of questions is divided into three levels: 1 (easy), 2 (medium), and 3 (difficult), requiring evaluators to score each question. In the research (*Wang et al., 2018*), evaluators are asked to mark each question as "true/false" for fluency (coherent and grammatically correct), question-answer relevance, and machine generation. *Steuer et al. (2021)* invited three trained education experts to evaluate question quality based on the annotation scheme developed by *Horbach et al. (2020)*, which consists of 6 binary indicators, 2 ternary indicators, and 1 quinary indicator, covering various aspects such as understandability, content relevance, naturalness, educational value, and cognitive levels of the questions. Obviously, manual evaluation is time-consuming and labor-intensive, requires high professional knowledge and experience from evaluators, and due to the subjectivity of the evaluation method, it is difficult to unify standards, leading to inconsistent evaluation results and limited widespread application.

DISCUSSIONS

As a branch of QG in the field of education, QGEd has already established a considerable research foundation and progress. With the continuous advancement of technology, especially breakthroughs in the fields of artificial intelligence and natural language processing, QGEd is gradually revealing its immense potential in educational innovation. For instance, the integration of deep learning models like transformers has enabled the generation of more complex and contextually relevant questions, moving beyond the limitations of traditional rule-based systems (*Vaswani et al., 2017*).

However, to meet the demands of the educational field, QGEd also faces numerous challenges. Among these, multimodal data processing stands out as a critical area. Educational questions should not be limited to text but should incorporate various forms such as images, audio, and video. While prior studies have explored visual question generation (VQG) (*Mulla & Gharpure, 2023*), these approaches often focus on general images rather than structured educational diagrams like mathematical geometric figures or flowcharts. Future research could explore how to adapt VQG techniques to better suit the specific requirements of educational content.

The controllability of fine-grained cognitive level and difficulty is another significant challenge. Questions should be designed to target specific cognitive abilities and difficulty levels to effectively facilitate learning. Although item response theory (IRT) has been widely used for question difficulty estimation (*Bi et al., 2024*), it often fails to capture the nuanced cognitive processes involved in answering questions. Recent studies have attempted to align question generation with established taxonomies like Bloom's Taxonomy (*Krathwohl, 2002*). For example, *Duong-Trung, Wang & Kravcik (2024)* proposed BloomLLM, which combines supervised fine-tuning with Bloom's Taxonomy to generate questions targeting specific cognitive levels.

The construction of specialized educational datasets is also crucial. Most existing datasets are either too generic or lack the necessary metadata for educational applications (*Al Faraby, Adiwijaya & Romadhony, 2024*). This limitation affects the quality and relevance of generated questions. Future work should focus on creating datasets that not only include questions and answers but also provide detailed annotations regarding educational objectives, cognitive levels, and difficulty. This would require collaboration between technologists and educational experts to ensure that the datasets meet the specific needs of educational settings.

Automatic evaluation technologies must be further developed to accurately assess the quality of generated questions. Traditional metrics like BLEU and ROUGE (*Papineni et al., 2002*; *Lin, 2004*) are limited in their ability to evaluate the educational value and cognitive level of questions. While newer metrics like BERTScore (*Zhang et al., 2020*) offer improvements by considering semantic similarity, they still fall short in capturing the educational significance of questions. Future research should focus on developing evaluation frameworks that incorporate educational theories and cognitive science principles.

System architecture design for QGEd systems also requires special attention. Unlike traditional QG systems, QGEd systems need to handle complex educational content and generate questions that align with specific teaching objectives. *Al Faraby, Adiwijaya & Romadhony (2024)* proposed a system architecture (as shown in Fig. 4) that includes key content identification, answer identification, and question construction. However, this

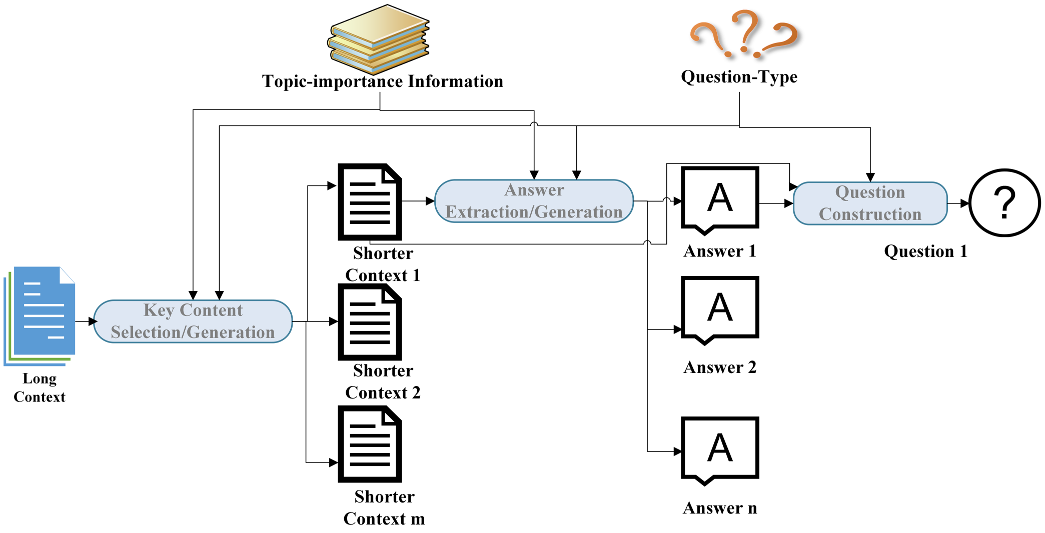
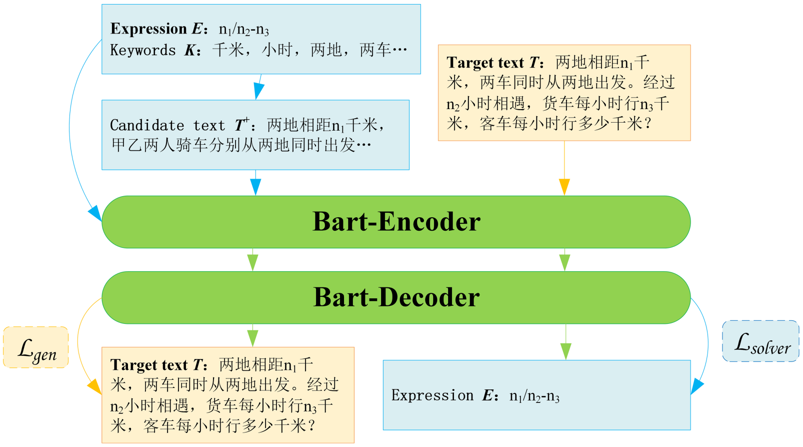


Figure 4 The structure diagram of the QGEd system for generating text questions based on text data.
Full-size DOI: 10.7717/peerj-cs.3203/fig-4



Expression E：n_1/n_2-n_3
Keywords K：千米，小时，两地，两车…

Candidate text T^{+}：两地相距n_1千米，甲乙两人骑车分别从两地同时出发…

Target text T：两地相距n_1千米，两车同时从两地出发。经过n_2小时相遇，货车每小时行n_3千米，客车每小时行多少千米？

Target text T：两地相距n_1千米，两车同时从两地出发。经过n_2小时相遇，货车每小时行n_3千米，客车每小时行多少千米？

Expression E：n_1/n_2-n_3

Figure 5 The model framework for generating math word questions.
Full-size DOI: 10.7717/peerj-cs.3203/fig-5

architecture primarily focuses on text-based questions. For STEM education, a more specialized architecture is needed that can handle complex problems involving multiple knowledge points and real-world scenarios. *SHEN (2022)* proposed a mathematical problem generation framework (as shown in Fig. 5. The diagram depicts a junior high math application problem where two cars travel towards each other and meet.千米: kilometer, 小时: hour, 两地: two places, 两车: two cars) that combines mathematical formulas with textual descriptions. This approach could serve as a foundation for developing more sophisticated QGEd systems for STEM subjects.

Looking ahead, with the continuous iteration and progress of general-purpose large language model technologies, cost-effective and efficient techniques such as fine-tuning of large models, prompt engineering, and retrieval-augmented generation (RAG) are likely to become hotspots in QGEd research. Meanwhile, the development of specialized datasets for QGEd should also be given high priority. Technologists should deepen their collaboration with educational researchers to further integrate QGEd technologies with educational theories, thereby ensuring that the generated questions can truly and accurately serve students' learning and cognitive development.

CONCLUSION

This review introduces the relationship between QGEd and QG, defines the three core objectives of QGEd—diversity, controllability, and focus on teaching content. We then analyze the technological evolution of QG, dividing it into heuristic rule-based and neural network technology stages, and discuss QGEd's evolution from 2015 to 2024. Based on this technological progression, we overview QGEd's current state across its three core objectives, datasets, and evaluation methods. We also identify pivotal challenges like the need for multimodal data processing, controllability of cognitive and difficulty levels, and specialized dataset development. To address these, we suggest future research directions, emphasizing automatic evaluation technology and system architecture design. Overall, this review offers valuable insights for researchers and practitioners, summarizing existing research and providing a foundation for QGEd's advancement.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The work is funded by the Science Foundation of Ministry of Education of China (No. 24YJA880007) and the National Natural Science Foundation of China (No. 62167007). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Science Foundation of Ministry of Education of China: 24YJA880007.
National Natural Science Foundation of China: 62167007.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Xiaohui Dong conceived and designed the experiments, performed the computation work, authored or reviewed drafts of the article, and approved the final draft.
- Xinyu Zhang conceived and designed the experiments, performed the experiments, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Zhengluo Li performed the experiments, analyzed the data, prepared figures and/or tables, and approved the final draft.

- Quanxin Hou analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Jixiang Xue analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Xiaoyi Li analyzed the data, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

This is a literature review.

REFERENCES

Al Faraby S, Adiwijaya A, Romadhony A. 2024. Review on neural question generation for education purposes. *International Journal of Artificial Intelligence in Education* **34(3)**:1008–1045 DOI 10.1007/s40593-023-00374-x.

Ali H, Chali Y, Hasan SA. 2010. Automatic question generation from sentences. In: Langlais P, Gagnon M, eds. *Actes de la 17e conférence sur le traitement automatique des langues naturelles. Articles courts, TALN 2010*. Montréal, Canada: ATALA, 213–218.

Ang BH, Gollapalli SD, Ng S-K. 2023. Socratic question generation: a novel dataset, models, and evaluation. In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023. May 2–6, 2023.* Stroudsburg: Association for Computational Linguistics, 147–165 DOI 10.18653/V1/2023.EACL-MAIN.12.

Artetxe M, Ruder S, Yogatama D. 2020. On the cross-lingual transferability of monolingual representations. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020. July 5–10, 2020.* Stroudsburg: Association for Computational Linguistics, 4623–4637 DOI 10.18653/V1/2020.ACL-MAIN.421.

Babakhani P, Lommatzsch A, Brodt T, Sacker D, Sivrikaya F, Albayrak S. 2024. Opinerium: subjective question generation using large language models. *IEEE Access* **12(1)**:66085–66099 DOI 10.1109/ACCESS.2024.3398553.

Banerjee S, Lavie A. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: Goldstein J, Lavie A, Lin C-Y, Voss CR, eds. *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*. Stroudsburg: Association for Computational Linguistics, 65–72.

Bao H, Dong L, Wei F, Wang W, Yang N, Liu X, Wang Y, Gao J, Piao S, Zhou M, Hon H-W. 2020. UniLMv2: pseudo-masked language models for unified language model pre-training, virtual event. In: *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13–18 July 2020, Virtual Event*. Cambridge: Proceedings of Machine Learning Research, 13–18.

Bi S, Liu J, Miao Z, Min Q. 2024. Difficulty-controllable question generation over knowledge graphs: a counterfactual reasoning approach. *Information Processing & Management* **61(4)**:103721 DOI 10.1016/J.IPM.2024.103721.

Bitew SK, Hadifar A, Sterckx L, Deleu J, Develder C, Demeester T. 2024. Learning to reuse distractors to support multiple-choice question generation in education. *IEEE Transactions on Learning Technologies* **17**:375–390 DOI 10.1109/TLT.2022.3226523.

Bloom BS. 1956. *Taxonomy of educational objectives: the classification of educational goals*. New York: Longman Green.

Cao S, Wang L. 2021. Controllable open-ended question generation with a new question type ontology. In: Zong C, Xia F, Li W, Navigli R, eds. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1–6, 2021*. Stroudsburg: Association for Computational Linguistics, 6424–6439.

Ch DR, Saha SK. 2023. Generation of multiple-choice questions from textbook contents of school-level subjects. *IEEE Transactions on Learning Technologies* **16(1)**:40–52 DOI 10.1109/TLT.2022.3224232.

Chai Z, Wan X. 2020. Learning to ask more: semi-autoregressive sequential question generation under dual-graph interaction. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 225–237.

Chali Y, Baghaee T. 2018. Automatic opinion question generation. In: Krahmer E, Gatt A, Goudbeek M, eds. *Proceedings of the 11th International Conference on Natural Language Generation, Tilburg University, The Netherlands, November 5–8, 2018*. Stroudsburg: Association for Computational Linguistics, 152–158.

Chan Y-H, Fan Y-C. 2019. A recurrent BERT-based model for question generation. In: Fisch A, Talmor A, Jia R, Seo M, Choi E, Chen D, eds. *Proceedings of the 2nd Workshop on Machine Reading for Question Answering, MRQA@EMNLP 2019, Hong Kong, China, November 4, 2019*. Stroudsburg: Association for Computational Linguistics, 154–162.

Chen G, Yang J, Gasevic D. 2019. A comparative study on question-worthy sentence selection strategies for educational question generation. In: Isotani S, Millán E, Ogan A, Hastings PM, McLaren BM, Luckin R, eds. *Artificial Intelligence in Education - 20th International Conference, AIED 2019, Chicago, IL, USA, June 25–29, 2019, Proceedings, Part I: Lecture Notes in Computer Science*. Berlin: Springer, 59–70.

Chen G, Yang J, Hauff C, Houben G-J. 2018. LearningQ: a large-scale dataset for educational question generation. In: *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018. June 25–28, 2018.* Stanford, California, USA, AAAI Press, 481–490.

Cheng Y, Li S, Liu B, Zhao R, Li S, Lin C, Zheng Y. 2021. Guiding the growth: difficulty-controllable question generation through step-by-step rewriting. In: Zong C, Xia F, Li W, Navigli R, eds. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021 (Volume 1: Long Papers), Virtual Event, August 1–6, 2021.* Stroudsburg: Association for Computational Linguistics, 5968–5978.

Chomphooyod P, Suchato A, Tuaycharoen N, Punyabukkana P. 2023. English grammar multiple-choice question generation using text-to-text transfer transformer. *Computers and Education: Artificial Intelligence* **5**:100158 DOI 10.1016/j.caeai.2023.100158.

Clark P, Cowhey I, Etzioni O, Khot T, Sabharwal A, Schoenick C, Tafjord O. 2018. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. ArXiv DOI 10.48550/ArXiv.1803.05457.

Cui S, Bao X, Zu X, Guo Y, Zhao Z, Zhang J, Chen H. 2021. OneStop QAMaker: extract question-answer pairs from text in a one-stop approach. ArXiv DOI 10.48550/ArXiv.2102.12128.

Cui Y, Liu T, Che W, Xiao L, Chen Z, Ma W, Wang S, Hu G. 2019. A span-extraction dataset for chinese machine reading comprehension. In: Inui K, Jiang J, Ng V, Wan X, eds. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th*

International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3–7, 2019. Stroudsburg: Association for Computational Linguistics, 5882–5888.

Dhole KD, Manning CD. 2020. Syn-QG: syntactic and shallow semantic rules for question generation. In: Jurafsky D, Chai J, Schluter N, Tetreault JR, eds. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*. Stroudsburg: Association for Computational Linguistics, 752–765.

Dong L, Yang N, Wang W, Wei F, Liu X, Wang Y, Gao J, Zhou M, Hon H-W. 2019. Unified language model pre-training for natural language understanding and generation. In: Wallach HM, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox EB, Garnett R, eds. *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8–14, 2019, Vancouver, BC, Canada*. Stroudsburg: Association for Computational Linguistics, 13042–13054.

Du X, Cardie C. 2017. Identifying where to focus in reading comprehension for neural question generation. In: Palmer M, Hwa R, Riedel S, eds. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9–11, 2017*. Stroudsburg: Association for Computational Linguistics.

Du X, Cardie C. 2018. Harvesting paragraph-level question-answer pairs from wikipedia. In: Gurevych I, Miyao Y, eds. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15–20, 2018, Volume 1: Long Papers*. Stroudsburg: Association for Computational Linguistics.

Du X, Shao J, Cardie C. 2017. Learning to ask: neural question generation for reading comprehension. In: Barzilay R, Kan M-Y, eds. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30–August 4, Volume 1: Long Papers*. Stroudsburg: Association for Computational Linguistics, 1342–1352.

Duong-Trung N, Wang X, Kravcik M. 2024. BloomLLM: large language models based question generation combining supervised fine-tuning and bloom's taxonomy. In: Ferreira Mello R, Rummel N, Jivet I, Pishtari G, Ruipérez Valiente JA, eds. *Technology enhanced learning for inclusive and equitable quality education. EC-TEL 2024: Lecture notes in computer science*. Vol. 15160. Cham: Springer.

Fabbri AR, Ng P, Wang Z, Nallapati R, Xiang B. 2020. Template-based question generation from retrieved sentences for improved unsupervised question answering. In: Jurafsky D, Chai J, Schluter N, Tetreault JR, eds. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5–10, 2020*. Stroudsburg: Association for Computational Linguistics, 4508–4513.

Fan A, Jernite Y, Perez E, Grangier D, Weston J, Auli M. 2019. ELI5: long form question answering. In: Korhonen A, Traum DR, Màrquez L, eds. *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28–August 2, 2019, Volume 1: Long Papers*. Stroudsburg: Association for Computational Linguistics, 3558–3567.

Fan Z, Wei Z, Wang S, Liu Y, Huang X. 2018. A Reinforcement learning framework for natural question generation using bi-discriminators. In: Bender EM, Derczynski L, Isabelle P, eds. *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20–26, 2018*. Stroudsburg: Association for Computational Linguistics, 1763–1774.

Flor M, Riordan B. 2018. A semantic role-based approach to open-domain automatic question generation. In: Tetreault JR, Burstein J, Kochmar E, Leacock C, Yannakoudakis H, eds. *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational*

Applications@NAACL-HLT 2018, New Orleans, LA, USA, June 5, 2018. Stroudsburg: Association for Computational Linguistics, 254–263.

Gao Y, Bing L, Chen W, Lyu MR, King I. 2019. Difficulty Controllable Generation of Reading Comprehension Questions. In: Kraus S, ed. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10–16, 2019*. International Joint Conferences on Artificial Intelligence, 4968–4974.

Gou Q, Xia Z, Yu B, Yu H, Huang F, Li Y, Nguyen C-T. 2023. Diversify question generation with retrieval-augmented style transfer. In: Bouamor H, Pino J, Bali K, eds. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6–10, 2023*. Stroudsburg: Association for Computational Linguistics, 1677–1690.

Graesser AC, Person NK. 1994. Question asking during tutoring. *American Educational Research Journal* **31(1)**:104–137 DOI 10.3102/00028312031001104.

Guo S, Zhang J, Ke X, Li C, Chen H. 2023. Diversifying question generation over knowledge base via external natural questions. In: Calzolari N, Kan M-Y, Hoste V, Lenci A, Sakti S, Xue N, eds. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20–25 May, 2024, Torino, Italy*. Paris, France: ELRA and ICCL, 5096–5108.

Hadifar A, Bitew SK, Deleu J, Develder C, Demeester T. 2023. EduQG: a multi-format multiple-choice dataset for the educational domain. *IEEE Access* **11(1)**:20885–20896 DOI 10.1109/ACCESS.2023.3248790.

He W, Liu K, Liu J, Lyu Y, Zhao S, Xiao X, Liu Y, Wang Y, Wu H, She Q, Liu X, Wu T, Wang H. 2018. DuReader: a Chinese machine reading comprehension dataset from real-world applications. In: Choi E, Seo M, Chen D, Jia R, Berant J, eds. *Proceedings of the Workshop on Machine Reading for Question Answering@ACL 2018, Melbourne, Australia, July 19, 2018*. Stroudsburg: Association for Computational Linguistics, 37–46.

Heilman M, Smith NA. 2010. Good question! statistical ranking for question generation. In: *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 2–4, 2010, Los Angeles, California, USA*. Stroudsburg: The Association for Computational Linguistics, 609–617.

Horbach A, Aldabe I, Bexte M, de LOL, Maritxalar M, Calzolari N, Béchet F, Blache P, Choukri K, Cieri C, Declerck T, Goggi S, Isahara H, Maegaard B, Mariani J, Mazo H, Moreno A. 2020. Linguistic appropriateness and pedagogic usefulness of reading comprehension questions. In: Odijk J, Piperidis S, eds. *Proceedings of the 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11–16, 2020*. Paris, France: European Language Resources Association, 1753–1762.

Huang L, Bras RL, Bhagavatula C, Choi Y. 2019. Cosmos QA: machine reading comprehension with contextual commonsense reasoning. In: Inui K, Jiang J, Ng V, Wan X, eds. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3–7, 2019*. Stroudsburg: Association for Computational Linguistics, 2391–2401.

Huang Y, He L. 2016. Automatic generation of short answer questions for reading comprehension assessment. *Natural Language Engineering* **22(3)**:457–489 DOI 10.1017/S1351324915000455.

Jauhar SK, Turney P, Hovy E. 2016. TabMCQ: a dataset of general knowledge tables and multiple-choice questions. ArXiv DOI 10.48550/ArXiv.1602.03960.

Jia X, Zhou W, Sun X, Wu Y. 2021. EQG-RACE: examination-type question generation. In: *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference* *on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2–9, 2021*. Washington DC: AAAI Press.

Kim Y, Lee H, Shin J, Jung K. 2019. Improving neural question generation using answer separation. In: *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 – February 1, 2019*. Washington DC: AAAI Press, 6602–6609.

Ko W-J, Chen T-Y, Huang Y, Durrett G, Li JJ. 2020. Inquisitive question generation for high level text comprehension. In: Webber B, Cohn T, He Y, Liu Y, eds. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16–20, 2020*. Stroudsburg: Association for Computational Linguistics, 6544–6555.

Koncel-Kedziorski R, Roy S, Amini A, Kushman N, Hajishirzi H. 2016. MAWPS: a math word problem repository. In: Knight K, Nenkova A, Rambow O, eds. *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12–17, 2016*. Stroudsburg: The Association for Computational Linguistics, 1152–1157.

Krathwohl DR. 2002. A revision of Bloom's taxonomy: an overview. *Theory Into Practice* **41**:212–218 DOI 10.1207/s15430421tip4104_2.

Kumar G, Banchs RE, D'Haro LF. 2015. Automatic fill-the-blank question generator for student self-assessment. In: *2015 IEEE Frontiers in Education Conference, FIE 2015, El Paso, TX, USA, October 21–24, 2015*. Piscataway: IEEE Computer Society, 1–3.

Kumar V, Joshi N, Mukherjee A, Ramakrishnan G, Jyothi P. 2019. Cross-lingual training for automatic question generation. In: Korhonen A, Traum DR, Màrquez L, eds. *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28–August 2, 2019, Volume 1: Long Papers*. Stroudsburg: Association for Computational Linguistics, 4863–4872.

Kunichika H, Katayama T, Hirashima T, Takeuchi A. 2004. Automated question generation methods for intelligent English learning systems and its evaluation. *Available at https://api.semanticscholar.org/CorpusID:1536529*.

Kurdi G, Leo J, Parsia B, Sattler U, Al-Emari S. 2020. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education* **30(1)**:121–204 DOI 10.1007/S40593-019-00186-Y.

Kwiatkowski T, Palomaki J, Redfield O, Collins M, Parikh AP, Alberti C, Epstein D, Polosukhin I, Devlin J, Lee K, Toutanova K, Jones L, Kelcey M, Chang M-W, Dai AM, Uszkoreit J, Le Q, Petrov S. 2019. Natural Questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics* **7(15)**:452–466 DOI 10.1162/TACL_A_00276.

Lai G, Xie Q, Liu H, Yang Y, Hovy EH. 2017. RACE: large-scale reading comprehension dataset from examinations. In: Palmer M, Hwa R, Riedel S, eds. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9–11, 2017*. Stroudsburg: Association for Computational Linguistics, 785–794.

Le N-T, Kojiri T, Pinkwart N. 2014. Automatic question generation for educational applications - the state of art. In: Do TV, Thi HAL, Nguyen NT, eds. *Advanced Computational Methods for Knowledge Engineering - Proceedings of the 2nd International Conference on Computer Science, Applied Mathematics and Applications, ICCSAMA 2014, 8–9 May, 2014, Budapest, Hungary: Advances in Intelligent Systems and Computing*. Berlin: Springer, 325–338.

Lee U, Jung H, Jeon Y, Sohn Y, Hwang W, Moon J, Kim H. 2024. Few-shot is enough: exploring ChatGPT prompt engineering method for automatic question generation in english education. *Education and Information Technologies* **29(9)**:11483–11515 DOI 10.1007/s10639-023-12249-8.

Lee DB, Lee S, Jeong WT, Kim D, Hwang SJ. 2020. Generating diverse and consistent QA pairs from contexts with information-maximizing hierarchical conditional VAEs. In: Jurafsky D, Chai J, Schluter N, Tetreault JR, eds. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5–10, 2020*. Stroudsburg, PA: Association for Computational Linguistics, 208–224.

Lee C-H, Wang S-M, Chang H-C, Lee H. 2018. ODSQA: open-domain spoken question answering dataset. In: *IEEE Spoken Language Technology Workshop, SLT 2018, Athens, Greece, December 18–21, 2018*. Piscataway: IEEE, 949–956.

Lewis PSH, Oguz B, Rinott R, Riedel S, Schwenk H. 2020. MLQA: evaluating cross-lingual extractive question answering. In: Jurafsky D, Chai J, Schluter N, Tetreault JR, eds. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5–10, 2020*. Stroudsburg: Association for Computational Linguistics, 7315–7330.

Liang Y, Li J, Yin J. 2019. A new multi-choice reading comprehension dataset for curriculum learning. In: Lee WS, Suzuki T, eds. *Proceedings of The 11th Asian Conference on Machine Learning, ACML 2019, 17–19 November 2019, Nagoya, Japan*. Cambridge: Proceedings of Machine Learning Research, 742–757.

Lin C-Y. 2004. ROUGE: a package for automatic evaluation of summaries. In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, 74–81.

Lindberg D, Popowich F, Nesbit JC, Winne PH. 2013. Generating natural language questions to support learning on-line. In: Gatt A, Saggion H, eds. *ENLG 2013 - Proceedings of the 14th European Workshop on Natural Language Generation, August 8-9, 2013, Sofia, Bulgaria*. Stroudsburg: The Association for Computer Linguistics, 105–114.

Liu T, Wei B, Chang B, Sui Z. 2017. Large-scale simple question generation by template-based Seq2seq learning. In: Huang X, Jiang J, Zhao D, Feng Y, Hong Y, eds. *Proceedings of the Natural Language Processing and Chinese Computing - 6th CCF International Conference, NLPCC 2017, Dalian, China, November 8–12, 2017: Lecture Notes in Computer Science*. Berlin: Springer, 75–87.

Liu B, Wei H, Niu D, Chen H, He Y. 2020. Asking questions the human way: scalable question-answer generation from text corpus. In: Huang Y, King I, Liu T-Y, van SM, eds. *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20–24, 2020*. New York: ACM, 2032–2043.

Liu M, Zhang J. 2022. Chinese neural question generation: augmenting knowledge into multiple neural encoders. *Applied Sciences* **12(3)**:1032 DOI 10.3390/app12031032.

Liu B, Zhao M, Niu D, Lai K, He Y, Wei H, Xu Y. 2019. Learning to generate questions by learningwhat not to generate. In: *The World Wide Web Conference*. San Francisco: ACM, 1106–1118 DOI 10.1145/3308558.3313737.

Ma X, Zhu Q, Zhou Y, Li X, Wu D. 2020. Asking complex questions with multi-hop answer-focused reasoning. ArXiv DOI 10.48550/ArXiv.2009.07402.

Maity S, Deroy A, Sarkar S. 2025. Leveraging in-context learning and retrieval-augmented generation for automatic question generation in educational domains. ArXiv DOI 10.48550/ArXiv.2501.17397.

Matsumori S, Okuoka K, Shibata R, Inoue M, Fukuchi Y, Imai M. 2023. Mask and cloze: automatic open cloze question generation using a masked language model. *IEEE Access* **11(4)**:9835–9850 DOI 10.1109/ACCESS.2023.3239005.

Mihaylov T, Clark P, Khot T, Sabharwal A. 2018. Can a suit of armor conduct electricity? A new dataset for open book question answering. In: Riloff E, Chiang D, Hockenmaier J, Tsujii J, eds. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*. Stroudsburg: Association for Computational Linguistics, 2381–2391.

Mitkov R, Ha LA. 2003. Computer-aided generation of multiple-choice tests. In: *Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing*, 17–22.

Moon H, Lee J, Eo S, Park C, Seo J, Lim H-S. 2024. Generative interpretation: toward human-like evaluation for educational question-answer pair generation. In: *Findings of the Association for Computational Linguistics: EACL 2024*, 2185–2196.

Mostow J, Chen W. 2009. Generating instruction automatically for the reading strategy of self-questioning. In: *Artificial Intelligence in Education*. Amsterdam: IOS Press, 465–472.

Mulla N, Gharpure P. 2023. Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications. *Progress in Artificial Intelligence* **12(1)**:1–32 DOI 10.1007/S13748-023-00295-9.

Murakhovs'ka L, Wu C-S, Laban P, Niu T, Liu W, Xiong C. 2021. MixQG: neural question generation with mixed answer types. In: Carpuat M, de Marneffe M-C, Ruíz IVM, eds. *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10–15, 2022*. Stroudsburg, PA: Association for Computational Linguistics, 1486–1497.

Olney AM, Graesser AC, Person NK. 2012. Question generation from concept maps. *Dialogue & Discourse* **3(2)**:75–99 DOI 10.5087/dad.2012.204.

Olney AM Jr, Pavlik PI, Maass JK. 2017. Improving reading comprehension with automatically generated cloze item practice. In: André E, Baker RS, Hu X, Rodrigo MMT, du Boulay B, eds. *Artificial Intelligence in Education - 18th International Conference, AIED 2017, Wuhan, China, June 28 - July 1, 2017, Proceedings*. Berlin: Springer, 262–273.

Pal A, Umapathi LK, Sankarasubbu M. 2022. MedMCQA: a large-scale multi-subject multi-choice dataset for medical domain question answering. In: Flores G, Chen GH, Pollard TJ, Ho JC, Naumann T, eds. *Conference on Health, Inference, and Learning, CHIL 2022, 7–8 April 2022, Virtual Event*. Cambridge: Proceedings of Machine Learning Research, 248–260.

Pan Y, Hu B, Chen Q, Xiang Y, Wang X. 2020. Learning to generate diverse questions from keywords. In: *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*. Piscataway: IEEE, 8224–8228.

Papineni K, Roukos S, Ward T, Zhu W-J. 2002. Bleu: a method for automatic evaluation of machine translation. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6–12, 2002, Philadelphia, PA, USA*. Stroudsburg: ACL, 311–318.

Park J, Cho H, Lee S. 2018. Automatic generation of multiple-choice fill-in-the-blank question using document embedding. In: Rosé CP, Maldonado RM, Hoppe HU, Luckin R, Mavrikis M, Porayska-Pomsta K, McLaren BM, du Boulay B, eds. *Artificial Intelligence in Education - 19th International Conference, AIED 2018, London, UK, June 27–30, 2018, Proceedings, Part II: Lecture Notes in Computer Science*. Berlin: Springer, 261–265.

Popović M. 2015. chrF: character n-gram F-score for automatic MT evaluation. In: *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Lisbon, Portugal: Association for Computational Linguistics, 392–395 DOI 10.18653/v1/W15-3049.

Priya TJ, Priya KPS, Jenneyl LR, Uma KV. 2022. Automatic question generation from video. In: Das AK, Nayak J, Naik B, Vimal S, Pelusi D, eds. *Computational intelligence in pattern* *recognition: Lecture notes in networks and systems*. Singapore: Springer Nature Singapore, 366–372.

Prokudin A, Sychev O, Denisov M. 2023. Learning problem generator for introductory programming courses. *Software Impacts* **17(1)**:100519 DOI 10.1016/J.SIMPA.2023.100519.

Qi W, Yan Y, Gong Y, Liu D, Duan N, Chen J, Zhang R, Zhou M. 2020. ProphetNet: predicting future N-gram for sequence-to-sequence pre-training. In: Cohn T, He Y, Liu Y, eds. *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16–20 November 2020*. Stroudsburg: Association for Computational Linguistics, 2401–2410.

Qu F, Jia X, Wu Y. 2021. Asking questions like educational experts: automatically generating question-answer pairs on real-world examination data. In: Moens M-F, Huang X, Specia L, Yih SW, eds. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7–11 November, 2021*. Stroudsburg: Association for Computational Linguistics, 2583–2593.

Rajpurkar P, Zhang J, Lopyrev K, Liang P. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In: Su J, Carreras X, Duh K, eds. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1–4, 2016*. Stroudsburg: The Association for Computational Linguistics, 2383–2392.

Rao S, Daumé H III. 2019. Answer-based adversarial training for generating clarification questions. In: Burstein J, Doran C, Solorio T, eds. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2–7, 2019, Volume 1 (Long and Short Papers)*. Stroudsburg: Association for Computational Linguistics, 143–155.

Reddy S, Chen D, Manning CD. 2019. CoQA: a conversational question answering challenge. *Transactions of the Association for Computational Linguistics* **7**:249–266 DOI 10.1162/TACL_A_00266.

Reddy S, Raghu D, Khapra MM, Joshi S. 2017. Generating natural language question-answer pairs from a knowledge graph using a RNN based question generation model. In: Lapata M, Blunsom P, Koller A, eds. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3–7, 2017, Volume 1: Long Papers*. Stroudsburg: Association for Computational Linguistics, 376–385.

Sang EFTK, Meulder FD. 2003. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. In: Daelemans W, Osborne M, eds. *Proceedings of the Seventh Conference on Natural Language Learning, CoNLL 2003, Held in cooperation with HLT-NAACL 2003, Edmonton, Canada, May 31–June 1, 2003*. Stroudsburg: Association for Computational Linguistics, 142–147.

Scialom T, Piwowarski B, Staiano J. 2019. Self-attention architectures for answer-agnostic neural question generation. In: Korhonen A, Traum DR, Màrquez L, eds. *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28 - August 2, 2019, Volume 1: Long Papers*. Stroudsburg: Association for Computational Linguistics, 6027–6032.

SHEN Y. 2022. Automatic solving and text generation of math word problems. PhD thesis. East China Normal University DOI 10.27149/d.cnki.ghdsu.2022.000781.

Shinoda K, Aizawa A. 2020. Variational question-answer pair generation for machine reading comprehension. ArXiv DOI 10.48550/ArXiv.2004.03238.

Shoaib M, Husnain G, Sayed N, Ghadi YY, Alajmi M, Qahmash A. 2025. Automated generation of multiple-choice questions for computer science education using conditional generative adversarial networks. *IEEE Access* **13(3)**:16697–16715 DOI 10.1109/ACCESS.2025.3530474.

Song L, Wang Z, Hamza W, Zhang Y, Gildea D. 2018. Leveraging context information for natural question generation. In: Walker MA, Ji H, Stent A, eds. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1–6, 2018, Volume 2 (Short Papers)*. Stroudsburg: Association for Computational Linguistics, 569–574.

Srivastava M, Goodman N. 2021. Question generation for adaptive education. In: Zong C, Xia F, Li W, Navigli R, eds. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Stroudsburg: Association for Computational Linguistics, 692–701.

Stasaski K, Rathod M, Tu T, Xiao Y, Hearst MA. 2021. Automatically generating cause-and-effect questions from passages. In: Burstein J, Horbach A, Kochmar E, Laarmann-Quante R, Leacock C, Madnani N, Pilán I, Yannakoudakis H, Zesch T, eds. *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications, BEA@EACL, Online, April 20, 2021*. Stroudsburg: Association for Computational Linguistics, 158–170.

Steuer T, Filighera A, Meuser T, Rensing C. 2021. I do not understand what I cannot define: automatic question generation with pedagogically-driven content selection. ArXiv DOI 10.48550/ArXiv.2110.04123.

Steuer T, Filighera A, Rensing C. 2020. Remember the facts? Investigating answer-aware neural question generation for text comprehension. In: Bittencourt II, Cukurova M, Muldner K, Luckin R, Millán E, eds. *Artificial Intelligence in Education - 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part I: Lecture Notes in Computer Science*. Berlin: Springer, 512–523.

Sun X, Liu J, Lyu Y, He W, Ma Y, Wang S. 2018. Answer-focused and position-aware neural question generation. In: Riloff E, Chiang D, Hockenmaier J, Tsujii J, eds. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 3930–3939.

Tang D, Duan N, Yan Z, Zhang Z, Sun Y, Liu S, Lv Y, Zhou M. 2018. Learning to collaborate for question answering and asking. In: Walker MA, Ji H, Stent A, eds. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1–6, 2018, Volume 1 (Long Papers)*. Stroudsburg: Association for Computational Linguistics, 1564–1574.

Teruyoshi G, Tomikawa Y, Masaki UTO. 2024. Enhancing diversity in difficulty-controllable question generation for reading comprehension via extended T5. In: *International Conference on Computers in Education,* Manila, Philippines.

Tomikawa Y, Masaki UTO. 2024. Difficulty-controllable reading comprehension question generation considering the difficulty of reading passages. In: *International Conference on Computers in Education,* Manila, Philippines.

Tomikawa Y, Suzuki A, Uto M. 2024. Adaptive question-answer generation with difficulty control using item response theory and pretrained transformer models. *IEEE Transactions on Learning Technologies* **17(8)**:2186–2198 DOI 10.1109/TLT.2024.3491801.

Tonga JC, Clement B, Oudeyer P-Y. 2024. Automatic generation of question hints for mathematics problems using large language models in educational technology. ArXiv DOI 10.48550/ArXiv.2411.03495.

Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, Moher D, Peters MDJ, Horsley T, Weeks L, Hempel S, Akl EA, Chang C, McGowan J, Stewart L, Hartling L,

Aldcroft A, Wilson MG, Garritty C, Lewin S, Godfrey CM, Macdonald MT, Langlois EV, Soares-Weiser K, Moriarty J, Clifford T, Tunçalp Ö, Straus SE. 2018. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. *Annals of Internal Medicine* **169(7)**:467–473 DOI 10.7326/M18-0850.

Trischler A, Wang T, Yuan X, Harris J, Sordoni A, Bachman P, Suleman K, Blunsom P, Bordes A, Cho K, Cohen SB, Dyer C, Grefenstette E, Hermann KM, Rimell L, Weston J. 2017. NewsQA: a machine comprehension dataset. In: Yih S, ed. *Proceedings of the 2nd Workshop on Representation Learning for NLP, Rep4NLP@ACL 2017, Vancouver, Canada, August 3, 2017*. Stroudsburg: Association for Computational Linguistics, 191–200.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* **30**.

Wang Z, Lan AS, Nie W, Waters AE, Grimaldi PJ, Baraniuk RG. 2018. QG-net: a data-driven question generation model for educational content. In: Luckin R, Klemmer SR, Koedinger KR, eds. *Proceedings of the Fifth Annual ACM Conference on Learning at Scale, London, UK, June 26-28, 2018*. New York: ACM, 7:1–7:10.

Wang Y, Liu X, Shi S. 2017. Deep neural solver for math word problems. In: Palmer M, Hwa R, Riedel S, eds. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9–11, 2017*. Stroudsburg: Association for Computational Linguistics, 845–854.

Wang Z, Rao S, Zhang J, Qin Z, Tian G, Wang J. 2020a. Diversify question generation with continuous content selectors and question type modeling. In: Cohn T, He Y, Liu Y, eds. *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16–20 November 2020*. Stroudsburg: Association for Computational Linguistics, 2134–2143.

Wang W, Wei F, Dong L, Bao H, Yang N, Zhou M. 2020b. MiniLM: deep self-attention distillation for task-agnostic compression of pre-trained transformers. In: *Advances in Neural Information Processing Systems*. New York: Curran Associates, Inc, 5776–5788.

Wang T, Yuan X, Trischler A. 2017. A joint model for question answering and question generation. Arxiv DOI 10.48550/ArXiv.1706.01450.

Wang Y, Zheng J, Liu Q, Zhao Z, Xiao J, Zhuang Y. 2019. Weak supervision enhanced generative network for question generation. In: Kraus S, ed. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10–16, 2019*. California: International Joint Conferences on Artificial Intelligence.

Welbl J, Liu NF, Gardner M. 2017. Crowdsourcing multiple choice science questions. In: Derczynski L, Xu W, Ritter A, Baldwin T, eds. *Proceedings of the 3rd Workshop on Noisy User-generated Text, NUT@EMNLP 2017, Copenhagen, Denmark, September 7, 2017*. Stroudsburg, PA: Association for Computational Linguistics, 94–106.

Wolfe JH. 1976. Automatic question generation from text—an aid to independent study. In: Colman RW Jr, Lorton P, eds. *Proceedings of the ACM SIGCSE-SIGCUE technical symposium on Computer science and education, SIGCSE-SIGCUE 1976, Anaheim, CA, USA, February 12–13, 1976*. New York: ACM, 104–112.

Wu K, Zhu M, Zhu H, Zhang Y, Hong Y. 2021. Joint question type and penalty mechanism for question generation. *JCIP* **35**:110–119 DOI 10.3969/j.issn.1003-0077.2021.04.015.

Xiao D, Zhang H, Li Y-K, Sun Y, Tian H, Wu H, Wang H. 2020. ERNIE-GEN: an enhanced multi-flow pre-training and fine-tuning framework for natural language generation. In: Bessiere C, ed. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*. California: International Joint Conferences on Artificial Intelligence, 3997–4003.

Xie Q, Lai G, Dai Z, Hovy EH. 2018. Large-scale cloze test dataset created by teachers. In: Riloff E, Chiang D, Hockenmaier J, Tsujii J, eds. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*. Stroudsburg: Association for Computational Linguistics, 2344–2356.

Xu Y, Wang D, Yu M, Ritchie D, Yao B, Wu T, Zhang Z, Li TJ-J, Bradford N, Sun B, Hoang TB, Sang Y, Hou Y, Ma X, Yang D, Peng N, Yu Z, Warschauer M. 2022. Fantastic questions and where to find them: FairytaleQA - an authentic dataset for narrative comprehension. In: Muresan S, Nakov P, Villavicencio A, eds. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22–27, 2022*. Stroudsburg, PA: Association for Computational Linguistics, 447–460.

Yang Z, Hu J, Salakhutdinov R, Cohen WW. 2017. Semi-supervised QA with generative domain-adaptive nets. In: Barzilay R, Kan M-Y, eds. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*. Stroudsburg: Association for Computational Linguistics, 1040–1050.

Yang Z, Qi P, Zhang S, Bengio Y, Cohen WW, Salakhutdinov R, Manning CD. 2018. HotpotQA: a dataset for diverse, explainable multi-hop question answering. In: Riloff E, Chiang D, Hockenmaier J, Tsujii J, eds. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*. Stroudsburg: Association for Computational Linguistics, 2369–2380.

Yao B, Wang D, Wu T, Zhang Z, Li TJ-J, Yu M, Xu Y. 2022. It is AI's turn to ask humans a question: question-answer pair generation for children's story books. DOI 10.48550/arXiv.2109.03423.

Yao X, Zhang Y. 2010. Question generation with minimal recursion semantics. In: *Proceedings of QG2010: The Third Workshop on Question Generation*. Citeseer, 68–75.

Yu W, Jiang Z, Dong Y, Feng J. 2020. ReClor: a reading comprehension dataset requiring logical reasoning. In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26–30, 2020.*

Yuan X, Wang T, Gülçehre Ç, Sordoni A, Bachman P, Zhang S, Subramanian S, Trischler A, Blunsom P, Bordes A, Cho K, Cohen SB, Dyer C, Grefenstette E, Hermann KM, Rimell L, Weston J. 2017. Machine comprehension by text-to-text neural question generation. In: Yih S, ed. *Proceedings of the 2nd Workshop on Representation Learning for NLP, Rep4NLP@ACL 2017, Vancouver, Canada, August 3, 2017*. Stroudsburg, PA: Association for Computational Linguistics, 15–25.

Yulan SU, Yu H, Hongyu Z, Kaili W, Min Z. 2022. Adaptive optimization method of pre-trained language model for question generation. *JCIP* **36**:91–100 DOI 10.3969/j.issn.1003-0077.2022.03.011.

Zeng B, Pei F, Xu M, Ding M. 2022. Dual attention-based paragraph-level question generation. *JCIP* **36**:154–162 DOI 10.3969/j.issn.1003-0077.2022.08.019.

Zhang T, Cui Z, Leng J, Liu Y. 2021. CSFQGD: Chinese sentence fill-in-the-blank question generation dataset for examination. In: Shen W, Barthès J-PA, Luo J, Shi Y, Zhang J, eds. *24th IEEE International Conference on Computer Supported Cooperative Work in Design, CSCWD 2021, Dalian, China, May 5–7, 2021*. Piscataway: IEEE, 609–613.

Zhang R, Guo J, Chen L, Fan Y, Cheng X. 2022. A review on question generation from natural language text. *ACM Transactions on Information Systems* **40(1)**:1–43 DOI 10.1145/3468889.

Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. 2020. BERTScore: evaluating text generation with BERT. ArXiv DOI 10.48550/ArXiv.1904.09675.

Zhang J, Xie J, Qiu X. 2024. Controlling cloze-test question item difficulty with PLM-based surrogate models for IRT assessment. ArXiv DOI 10.48550/ArXiv.2403.01456.

Zhang Z, Zhu KQ. 2021. Diverse and specific clarification question generation with keywords. In: Leskovec J, Grobelnik M, Najork M, Tang J, Zia L, eds. *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19–23, 2021*. New York: ACM, 3501–3511.

Zhao Y, Ni X, Ding Y, Ke Q. 2018. Paragraph-level neural question generation with maxout pointer and gated self-attention networks. In: Riloff E, Chiang D, Hockenmaier J, Tsujii J, eds. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*. Stroudsburg, PA: Association for Computational Linguistics, 3901–3910.

Zheng H-T, Han J, Chen J-Y, Sangaiah AK. 2018. A novel framework for automatic Chinese question generation based on multi-feature neural network model. *Computer Science and Information Systems* **15(3)**:487–499 DOI 10.2298/CSIS171121018Z.

Zhou Q, Yang N, Wei F, Tan C, Bao H, Zhou M. 2017. Neural Question generation from text: a preliminary study. In: Huang X, Jiang J, Zhao D, Feng Y, Hong Y, eds. *Proceedings of the Natural Language Processing and Chinese Computing - 6th CCF International Conference, NLPCC 2017, Dalian, China, November 8–12, 2017: Lecture Notes in Computer Science*. Berlin: Springer, 662–671.

Zhou J, Zhou M, Gan J, Xu J. 2023. A question generation model with multi-stage temporal and semantic information enhancement. *Computer Engineering & Science* **45**:1847 DOI 10.3969/j.issn.1007-130X.2023.10.015.

Zou B, Li P, Pan L, Aw AT. 2022. Automatic true/false question generation for educational purpose. In: Kochmar E, Burstein J, Horbach A, Laarmann-Quante R, Madnani N, Tack A, Yaneva V, Yuan Z, Zesch T, eds. *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*. Stroudsburg, PA: Association for Computational Linguistics, 61–70.

Zyrianova M, Kalpakchi D, Boye J. 2023. EMBRACE: evaluation and modifications for boosting RACE. ArXiv DOI 10.48550/ArXiv.2305.08433.