Revista Brasileira de Informática na Educação – RBIE
Brazilian Journal of Computers in Education
(ISSN online: 2317-6121; print: 1414-5685)
https://sol.sbc.org.br/journals/index.php/rbie

Submission: 06/Jun/2025; 1st round notif.: 12/Aug/2025; New version: 11/Aug/2025; 2nd round notif.: 10/Nov/2025;
Camera ready: 11/Dec/2026; Edition review: 27/Feb/2026; Available online: 26/Fev/2026; Published: 26/Fev/2026;

The Last Decade of Automatic Question Generation: A Review of Techniques, Limitations, and Applications in Business Process Management Education

Guilherme Rego Rockembach
Universidade Federal do Rio Grande do Sul
ORCID: 0000-0002-5055-4808
guilherme.rockembach@ufrgs.br

Lucineia Heloisa Thom
Universidade Federal do Rio Grande do Sul
ORCID: 0000-0002-0620-9302
lucineia@inf.ufrgs.br

Abstract

Automatic Question Generation (AQG) is a research area that employs Natural Language Processing (NLP) techniques to automatically generate questions from textual content. Although it is still considered an emerging field, AQG has experienced significant growth in recent years, driven by advances in artificial intelligence, especially in deep learning and large language models, as well as by the increasing demand for scalable educational technologies. This article presents a Systematic Literature Review (SLR) focused on AQG research conducted over the last decade. The review aimed to identify and analyze the main computational approaches, practical applications, existing limitations, evaluation methods, and the degree of acceptance by education professionals. The SLR was carried out using major academic databases, resulting in the selection of 103 relevant studies, of which 90 are original research articles and 13 are literature reviews. The results show a gap in studies applying AQG to Business Process Management education, highlighting an area where further investigation is needed. The review also identifies a clear trend toward the adoption of Transformer-based models, which have significantly improved question generation quality. However, the analysis also reveals a lack of consensus regarding standardized evaluation metrics, particularly for automatic assessments, and a notable gap in studies that investigate how educational professionals perceive and accept questions generated by these systems. This highlights an important area for future research.

Keywords: *Automatic Question Generation; Automatic Assessment; Test Generation; Business Process Management; BPM.*

Cite as: Rockembach, G. R., & Thom L. H. (2026).The Last Decade of Automatic Question Generation: A Review of Techniques, Limitations, and Applications in Business Process Management Education. Revista Brasileira de Informática na Educação, 34, 219-251. https://doi.org/10.5753/rbie.2026.6093.

1 Introduction

Automatic Question Generation (AQG) refers to the use of algorithms and natural language processing (NLP) techniques to create assessment questions from texts (Hwang & Utami, 2024; Zimerman et al., 2024). Although still under development, this field has shown significant growth, driven by recent advances in artificial intelligence (AI) and available computational power. These advancements have enabled the creation of more sophisticated systems, increasing interest and research aimed at improving these technologies (Maity et al., 2024). In massive open online courses (MOOCs) and other distance learning (DL) platforms, this technology can be useful for assessing students' progress in a scalable and potentially personalized manner, minimizing the need for constant human tutor intervention, especially since it becomes humanly impossible to correct the large volume of questions manually. (Izo et al., 2022). When integrated into chatbots or intelligent tutors, AQG has the potential to provide continuous support, adjusting to each student's pace and comprehension level, which could contribute to more effective learning (Thüs et al., 2024).

In the corporate environment, AQG can also find applications, particularly in the training of new employees. By creating questionnaires that guide employees in learning new roles, this technology can enable a more autonomous and continuous training process (E. Li et al., 2020). This could eventually optimize supervisors' time and improve employee information retention. While these possibilities are promising, the practical application and actual impact still depend on factors such as the refinement of methods and user acceptance of the technologies. In this regard, AQG could become a strategic tool not only in DL and corporate training but also in areas such as adaptive learning and academic research, highlighting its potential in both educational and professional settings (Lafkiar & En Nahnahi, 2024).

In this article, however, our focus lies on the educational domain, more specifically on Business Process Management (BPM), a discipline that bridges academic and professional contexts. BPM provides a structured approach to aligning business operations with organizational goals, involving modeling, analysis, execution, and optimization of processes (Dumas et al., 2013; Weske et al., 2007). Despite its importance, the literature highlights a shortage of information technology tools explicitly designed for BPM education (Chow, 2021; Wiechetek et al., 2017). This gap limits the development of essential skills among students and professionals, reinforcing the need to explore new approaches.

In this context, AQG emerges as a potential contribution to BPM teaching, yet its role remains underexplored (Rockembach & Thom, 2024). While several reviews have mapped AQG techniques and applications in general education (Alshboul & Baksa-Varga, 2022; Gorgun & Bulut, 2024), there is still no systematic examination of how these advances relate to BPM, nor whether recent developments in generative AI (Zeghouani et al., 2024) have addressed the lack of dedicated tools.

Therefore, the problem addressed in this article is the absence of a clear understanding of how AQG can support BPM education and whether existing approaches effectively fill this gap. To address this problem, the article conducts a Systematic Literature Review (SLR) with a descriptive and critical perspective, aiming to: (i) map AQG approaches in general education, without restriction regarding educational level or discipline; (ii) analyze those applied to BPM; and (iii) identify current limitations and opportunities for future research. Based on these objectives,

we propose the following Research Questions (RQs):

- RQ1 - What approaches have been used in AQG?
- RQ2 - What approaches are focused on BPM education?
- RQ3 - What limitations do existing approaches present?
- RQ4 - How has Generative AI influenced the field of AQG?
- RQ5 - What metrics or criteria are used to evaluate the quality of generated questions?
- RQ6 - What resistance do experts have regarding the automatic generation of content questions, especially among education professionals?

To answer the RQs, the article is organized as follows: the Background and Related Work section presents the theoretical foundations and previous studies that contextualize the research; the Methodology section presents the search strategy and the criteria for selecting and refining the studies; the Results section presents the original studies and literature reviews found, along with the analysis and answers to the RQs; finally, the Conclusion section provides a summary of the identified state of the art, highlights potential gaps for future research and the possible limitations of this study.

2 Background and Related Work

In recent years, there has been a substantial increase in the use of computational techniques aimed at supporting the teaching and learning process, particularly with the application of AI and NLP in educational contexts (Reis et al., 2024). One of the most promising applications of these techniques is AQG, which has the potential to enhance both formative and summative assessment processes (Izo et al., 2022). In this context, NLP techniques have been applied to extract, structure, and generate questions from educational texts (Neto et al., 2024). The field of AQG has evolved rapidly over the past decade, especially with the introduction of the Transformer architecture (since 2017) and Large Language Models (LLMs), which have raised the technical level of text generation applications across different domains (da Silva et al., 2023), and have had a direct impact on AQG (Santi et al., 2022).

Various AQG techniques have been developed over the years, each expanding the scope of possible applications. Early approaches based on rules and templates enabled the structured creation of questions, particularly useful in educational contexts with well-defined domains (Kurdi et al., 2020). Subsequently, statistical and traditional machine learning methods provided greater adaptability, allowing linguistic features to be leveraged to enrich question generation (Alshboul & Baksa-Varga, 2022). The advancement to neural networks increased the naturalness and fluency of generated questions, opening possibilities for intelligent tutoring systems and learning support applications (Al Faraby et al., 2023). More recently, models based on LLMs and Transformers have excelled in generating diverse and contextually relevant questions, enhancing the use of AQG at scale across multiple domains (García-Méndez et al., 2024; Madri & Meruva, 2023).

AQG has been applied in practice across various disciplines. For example, in courses on Object-Oriented Programming, NLP techniques have been used to automatically generate questions from educational texts, enabling rapid and adaptive assessment (Neto et al., 2024). In online courses and MOOCs, automatic question generation supports learning evaluation and provides almost immediate feedback to students, covering a wide range of subjects (Izo et al., 2022). More recently, Transformer-based models have been explored in intelligent tutoring systems, enhancing the diversity and quality of generated questions and demonstrating potential to support personalized learning across different areas of knowledge (Santi et al., 2022).

Despite these advances, AQG still faces several challenges. Rule-based approaches have limited generalization power, being confined to very specific contexts (Zhang et al., 2021). Statistical and traditional machine learning methods rely heavily on manual linguistic resources, which limits portability across languages and domains (Das et al., 2021b). Neural networks require large annotated datasets and present interpretability and generation control issues (Mulla & Gharpure, 2023a). LLM-based models introduce new difficulties, such as potential training data bias, hallucination of information, and the absence of standardized metrics to assess question quality (Al Shuraiqi et al., 2024; Gorgun & Bulut, 2024). Additionally, pedagogical alignment remains a concern: even when questions are grammatically correct, they are not always aligned with intended learning objectives (Tuhpatussania et al., 2024). These challenges indicate that technical progress must be accompanied by advances in evaluation, data curation, and pedagogical integration.

In summary, AQG has evolved from rule-based methods to neural networks and LLMs, increasing the diversity and naturalness of generated questions. Despite challenges such as inconsistent evaluation and pedagogical alignment, these techniques offer great potential to support learning in various contexts. There is room for a systematic examination of how recent developments in generative AI may impact this field, as well as their specific applications in the teaching of disciplines such as BPM. This scenario highlights the relevance of a review that investigates the state of the art and identifies gaps to be explored.

3 Methodology

The SLR proposed in this article was based on the methodology proposed by Kitchenham and Charters (2007), widely adopted in software engineering. Following its guidelines, the process was organized into three main phases: protocol planning, execution (study search, selection, and exclusion), and finally data extraction, analysis, and synthesis of results.

3.1 Search strategy and work selection

The research began in May 2024, when the articles published up to that date were collected and their analysis initiated. In January 2025, the study was updated with the inclusion and analysis of publications released after the first collection, ensuring coverage of the last 10 years (2015–2024). This time frame was chosen to ensure both the relevance and timeliness of the analyzed works, considering that the field of AQG has evolved significantly over the past decade. Advances such as the development of deep learning techniques, the emergence of large language models, and the growing interest in educational technologies have greatly influenced research in this area.

Table 1: Search strings.

(”automatic question generation” OR ”aqg” OR ”automated question generation”) AND (”approaches” OR ”methods” OR ”strategies” OR ”techniques” OR ”proposes”) AND (”quality assessment” OR ”question evaluation” OR ”metrics” OR ”question quality” OR ”limitations”)
(”geração automática de perguntas” OR ”geração automatizada de perguntas”) AND (”abordagem” OR ”método” OR ”estratégia” OR ”técnica” OR ”proposta”) AND (”avaliação de qualidade” OR ”avaliação de perguntas” OR ”métricas” OR ”qualidade da pergunta” OR ”limitações”)

Therefore, selecting this period allows for the inclusion of the most representative and up-to-date contributions, as well as the identification of recent trends and research gaps. The selected databases for the search were ACM, Scopus, Springer, and IEEE, recognized for their reliability and broad coverage in the fields of computer science and education. Additionally, Google Scholar, SBC-OpenLib (SOL) and Periódicos CAPES were used to search for works written in Portuguese, as finding publications in Portuguese on the other mentioned platforms can be challenging.

To construct the search string, we initially selected one representative keyword for each area of interest (automatic question generation, methodological approaches, and quality assessment). Then, we progressively expanded the query by including synonyms and variations connected with the Boolean operator OR, aiming to broaden the coverage of potentially relevant studies. This iterative process, which resulted in the final string, can be seen in Table 1.

The string was tested both in English and in Portuguese, in order to capture works published in both languages. In parallel, we also tested combinations that included terms related to Business Process Management (e.g., ”BPM”, ”business processes”, ”process modeling”). However, whenever these terms were incorporated into the query, the results set was empty, indicating that no studies directly connecting AQG with BPM were indexed in the consulted databases.

The search in English resulted in 1,439 documents, distributed across the following databases: ACM (113), Scopus (646), Springer (669), IEEE (11), SOL (0) and Periódicos CAPES (0). In contrast, the search in Portuguese yielded 13 documents, all found in the Google Scholar database, while the other databases returned no results. In total, 1,452 documents were identified.

3.1.1 Exclusion and inclusion criteria

After identification, documents were screened according to exclusion criteria (ECs): duplicates, non-article documents, lack of full-text availability, languages other than Portuguese or English, or studies outside the research scope. Zotero (Corporation for Digital Scholarship, 2024) was used to organize the PDFs and eliminate duplicates, while Rayyan (Ouzzani et al., 2016) facilitated the systematic application of the ECs and later the inclusion criteria (ICs). The records that remained after removing duplicates and non-article entries, as well as the notes and screening decisions registered during the review in Rayyan, are available in the dataset exported from Rayyan and stored on Figshare (Figshare, 2025). Figure 1 illustrates the progressive exclusion process.

The 128 remaining documents were then evaluated against the inclusion criteria:

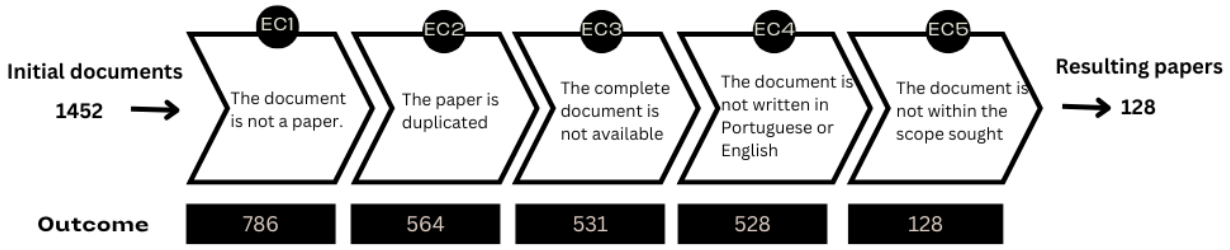

Figure 1: Application of ECs.

- (IC1) Methodology for automatic generation of assessment questions;
- (IC2) Evaluation of the quality of generated questions, including metrics;
- (IC3) Identification of limitations in the AQG field;
- (IC4) Assessment of experts' acceptance of automatically generated questions.

The evaluation sequence followed abstract → conclusion → methodology → full text (if needed). This process resulted in 103 included studies (13 reviews, 90 primary studies). The reviews were selected using the same criteria as in the original studies and were only separated later for data extraction, analysis, and reporting in the results.

3.1.2 Data extraction and tracking

Detailed information from the included studies was recorded in Google Sheets, including bibliographic metadata, methodological aspects, reported results, limitations, and other relevant content identified during reading. Labels in Rayyan were used to track which EC or IC justified each inclusion/exclusion and to mark the RQs addressed by each study. For instance, Gašpar et al., 2023 was included based on IC1 and IC2.

Review studies were analyzed separately from primary studies to allow a dedicated section summarizing what previous reviews covered and to highlight gaps not addressed by them. This process enabled mapping of all included studies to the six RQs defined in Section 1.

4 Results

In this section, the research results are presented. In particular, the information found in the reviewed studies is compiled, aiming to answer the RQs. The distribution concerning the timeline of the original works, i.e., those that are not literature reviews, are described in Figure 2. It is noticeable that there has been an increase in the number of publications in AQG starting from 2021, reaching its peak in 2024, with 39 papers on the topic, more than double the number from the previous year. This rise is likely due to the popularization of transformer-based models, as most of the recent works have adopted this approach, as shown in Table 3. Regarding the affiliation of the authors of these works, we can observe in Figure 3(a) that the main origin of the papers is the

Figure 2: Number of Publications per Year.

Asian continent, contributing 57 papers, with India being the country with the most publications, totaling 20 articles. The second continent with the most publications is Europe, contributing 24 papers on AQG. The only continent where no works were found is South America.

The knowledge areas of the questions generated from the approaches proposed in the primary studies were diverse, totaling thirteen knowledge areas. Figure 3(b) shows the distribution of these studies in relation to the application area. It can be observed that most studies focus on generating general knowledge questions, followed by those aimed at generating questions for language learning and Computer Science.

This section begins with a synthesis of the review papers on AQG found in the literature, as well as the gaps left by them, which justifies the existence of the current work. The remainder of the section is organized to address the RQs, and it concludes with a Summary of Results.

4.1 Review of reviews

In recent years, some SLRs have explored AQG, each offering significant insights into methodologies, applications, and challenges faced in the field. Among the SLRs conducted in the AQG domain, we find studies focused exclusively on investigating techniques for generating multiple-choice questions (MCQs), such as the works of Ch and Saha (2020) and Madri and Meruva (2023). Awalurahman and Budi (2024a) exclusively reviewed techniques for automatically generating distractors, one of the steps in generating MCQs. Al Shuraiqi et al. (2024), on the other hand, analyzed methodologies aimed at generating MCQs in a specific domain, the medical field.

Other SLRs papers have focused on surveying works that explore specific approaches to

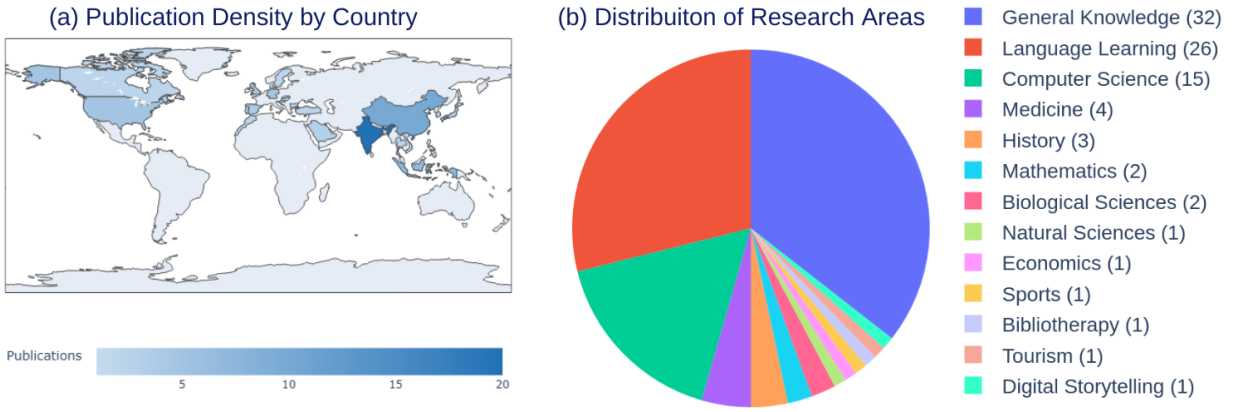

Figure 3: Distribution of works by authors' affiliations and application areas.

AQG. For instance, García-Méndez et al. (2024) reviewed works that exclusively used LLMs for automatic question generation, while Al Faraby et al. (2023) looked for works that generated questions using Artificial Neural Networks (ANNs), known as Neural Question Generation (NQG).

Some studies did not have restrictions regarding the domain or the type of question; however, they focused on a different time frame than the one intended in this SLR. The study by Kurdi et al. (2020) covered works published until 2019, that is, half of the current decade. Although with a limited time focus, the work by Kurdi et al. (2020), highlighted important elements for the field of AQG, such as the significant gap in the lack of standardized metrics to evaluate the quality of generated questions. Similarly to Kurdi et al. (2020), Alshboul and Baksa-Varga (2022) present a different time frame of publications than the one aimed for in the present work. The studies analyzed by Alshboul and Baksa-Varga (2022) were those published between 2017 and 2022.

The work by Gorgun and Bulut (2024) focused on reviewing evaluation methods and quality criteria in AQG, identifying a variety of approaches used by researchers. On the other hand, the study by Tuhpatussania et al. (2024) aimed to review AQG studies based on Indonesian texts, highlighting a lack of open datasets specific to this language.

Zhang et al. (2021) present a holistic view of question generation, proposing a comprehensive and systematic categorization of AQG tasks, aiming to answer how these tasks can be classified considering multiple factors. On the other hand, Das et al. (2021b) discuss existing methods and challenges associated with question generation from texts. Despite the important insights provided by both works, their approaches only address part of the research questions of this study and do not cover more recent advancements in the field, as shown in Figure 3. A more recent study by Mulla and Gharpure (2023a), expands on this perspective by identifying the same techniques discussed by Das et al. (2021b) for question generation from texts, in addition to including computer vision-based methods for question generation from images. However, even with this expansion, the work by Mulla and Gharpure (2023a) does not answer all the RQs of this study, highlighting the need for a more comprehensive approach to address the existing gaps in the literature.

The studies mentioned in this section provide a comprehensive overview of the state of the art in AQG; however, they do not fully address the objectives of the current study. By comparing these reviews, it becomes evident that gaps remain, particularly regarding: (i) coverage of the last

decade including the impact of generative AI, (ii) cross-domain applicability, and (iii) relevance to specific educational contexts such as BPM education. This analysis underscores the originality and necessity of the present SLR, which aims to provide a more complete, updated, and domain-oriented examination of AQG.

Table 2 summarizes the main characteristics of previous SLRs on AQG, highlighting their temporal scope, methodological criteria, thematic focus, and main limitations. This comparative view shows that although several reviews have advanced our understanding of AQG, they remain limited in at least one of three aspects: (i) restricted scope to specific techniques (e.g., MCQs, distractor generation, or neural models); (ii) narrow temporal coverage, often excluding recent works influenced by large language models; or (iii) lack of connection to specific educational domains such as BPM.

Table 2: Comparative overview of previous SLRs on AQG.

Studies	**Time frame**	**Focus / Methodological Scope**	**Main limitation**
Ch and Saha (2020); Madri and Meruva (2023)	Until 2020 / 2023	MCQ generation techniques	Limited to MCQ format
Awalurahman and Budi (2024a)	Until 2024	Distractor generation	Focused only on one AQG step
Al Shuraiqi et al. (2024)	Until 2024	MCQ in medical domain	Restricted to a single domain
García-Méndez et al. (2024)	Until 2024	LLM-based AQG	Narrow focus on one technique
Al Faraby et al. (2023)	Until 2023	Neural models (NQG)	Technique-specific review
Kurdi et al. (2020)	Until 2019	Broad AQG overview	Outdated temporal scope
Alshboul and Baksa-Varga (2022)	2017–2022	Broad AQG overview	Excludes recent advances in LLMs
Zhang et al. (2021); Das et al. (2021b); Mulla and Gharpure (2023a)	Until 2023	General AQG tasks and challenges	Partial coverage of RQs, limited to text-based AQG
Gorgun and Bulut (2024)	Until 2024	Evaluation methods	Focused only on assessment metrics
Tuhpatussania et al. (2024)	Until 2024	AQG in Indonesian texts	Language-specific limitation

As shown in Table 2, no previous review simultaneously covers the last decade, addresses the influence of recent advances such as LLMs, and considers implications for specific disciplines like BPM education. This highlights the originality and necessity of the present SLR, which

Table 3: Approaches adopted.

Approach	Publication period	Number of papers
Templates and Rules	2015, 2019, 2021, 2022, 2023, 2024	8
NLP techniques	2015, 2016, 2017, 2020, 2021, 2022, 2023, 2024	12
Transformers	2021, 2022, 2023, 2024	43
Ontology	2015, 2016, 2019, 2020, 2022, 2023	6
Deep Learning	2018, 2019, 2023, 2024	5
Combined techniques / Comparison between techniques	2021, 2022, 2023, 2024	16

provides a broader, updated, and domain-oriented analysis of AQG.

4.2 Answering RQ1 "What approaches have been used in AQG?"

When analyzing the 90 papers that contained primary studies, we found different types of approaches focused on AQG. To facilitate the understanding and discussion of the techniques used, we present the reviewed approaches grouped into seven categories. These categories were defined based on the type of question generation strategy adopted by the authors in each paper. Each category corresponds to a subsection of this text. The categories are: Template- and Rule-Based Approaches; Approaches Based on the Integration of NLP Techniques; Transformer-Based Approaches; Ontology-Based Approaches; Deep Learning-Based Approaches – Except Transformers; and Combined Architectures-Based Approaches and Comparative Analyses. A summary of the approaches adopted by the papers can be seen in Table 3.

4.2.1 Template- and Rule-Based Approaches

The most rudimentary yet efficient ways to generate questions automatically involve using prebuilt question models, known as templates, or question construction rules Keklik et al. (2019). Templates contain placeholder variables, which are filled in according to the context of the question to be generated. On the other hand, question construction rules define how the keywords extracted from a text (such as nouns or verbs) should be organized to form coherent questions. These methods are relatively simple to implement and computationally efficient, though often limited in coverage and linguistic variety.

Some approaches focus exclusively on predefined templates, organized according to sentence structures (e.g., cause, purpose, location) Keklik et al. (2019), or associated with entity types in structured data Shirude et al. (2015). Other works combine templates with external resources, such as domain knowledge bases Ruiz-Calleja et al. (2021), to increase domain specificity. A second line of research explores hybrid strategies that integrate template filling with machine learning

rules. For instance, Rodrigues et al. (2022) employ manually provided sentence–question pairs to identify recurrent patterns, progressively refining the rule set.

The sophistication of rule systems also varies. Wijanarko et al. (2021) combine TF-IDF keyword extraction with POS tagging, using Context-Free Grammar and Bloom's Taxonomy as a basis for categorization. Similarly, Gašpar et al. (2023) transform declarative sentences into interrogatives through Semantic Role Labeling, replacing arguments with question words according to their syntactic function.

Overall, template- and rule-based methods offer transparency and control in the generation process, but their reliance on handcrafted resources makes them less scalable and adaptable to open-domain contexts. From an educational perspective, their simplicity and interpretability can support teacher adoption and integration into classroom practices. However, their limited variety restricts the depth of assessment and poses challenges for promoting higher-order thinking skills. To facilitate comparison across studies, Table 4 summarizes representative contributions, highlighting their input data, techniques, and limitations.

Table 4: Representative template- and rule-based approaches for automatic question generation.

Study	**Data/Input**	**Technique**	**Highlights / Limitations**
Keklik et al. (2019)	Text sentences	Templates by sentence structure	Simple, efficient / limited variety
Shirude et al. (2015)	Structured tables	Entity-based templates	Domain-dependent
Ruiz-Calleja et al. (2021)	Domain knowledge base	Templates linked to KB	Strong domain grounding / low transferability
Rodrigues et al. (2022)	Sentence–question pairs	Rule learning from examples	Progressive refinement / requires annotated data
Wijanarko et al. (2021)	Unstructured text	TF-IDF, POS tagging, CFG, Bloom's Taxonomy	Rich categorization / complex pipeline
Gašpar et al. (2023)	Declarative sentences	SRL + linguistic rules	Semantic-driven / sensitive to parsing errors

4.2.2 Approaches Based on the Integration of Natural Language Processing Techniques

Automatic question generation systems can use a variety of NLP techniques to ensure the quality and relevance of the questions created, generating questions of quality comparable to those created by humans (Susanti et al. (2017), Chinkina et al. (2020) and Bhadouria et al. (2024)). A common practice in AQG works that use NLP techniques, as demonstrated by Mostow et al. (2016), Das and Majumder (2017), Jadhav et al. (2024) and Ch and Saha (2023), is the identification of parts of speech and the analysis of sentence syntactic structure.

Some works integrate multiple NLP modules. For instance, Larranaga et al. (2022) extract key concepts from educational texts and transform them into questions. Others construct detailed

syntactic trees to capture grammatical and semantic relations (Sewunetie and Kovacs (2024) and Narayana et al. (2024)). A representative case is Lui et al. (2015), who generate questions from Chinese texts by building syntactic parsing trees enriched with linguistic rules. Their system identifies categories such as nouns and verbs, as well as named entities, and applies specific rules to transform declarative sentences into causal, positional, or argumentative questions. This illustrates how parsing-based approaches combine structural and semantic information to increase the variety of generated questions.

Another research line focuses on selecting the most informative sentences for question generation. Odilinye et al. (2015), Le and Pinkwart (2015), and Shwe et al. (2024) adopt ranking algorithms based on the presence of key terms and semantic similarity (e.g., using WordNet) to align candidate sentences with learning objectives. Extending this process to multiple-choice questions, Das et al. (2021a) apply Rapid Automatic Keyword Extraction (RAKE, an unsupervised statistical method for extracting keywords from a text corpus) for keyword extraction, Named Entity Recognition (NER) for answer identification, and K-means clustering for distractor selection, ensuring semantic consistency across alternatives.

In general, NLP-based methods provide greater adaptability and linguistic coverage than template- or rule-based approaches. However, they often require complex pipelines, annotated resources, and significant computational power. From an educational standpoint, these methods enable richer and more context-aware question generation, which can improve alignment with learning materials. At the same time, their technical complexity and dependence on resources may limit accessibility for educators and institutions with fewer computational or linguistic resources. To organize these contributions, Table 5 summarizes key works, their data sources, techniques, and limitations.

Table 5: Representative NLP-based approaches for automatic question generation.

Study	**Data/Input**	**Techniques**	**Highlights / Limitations**
Lui et al. (2015)	Chinese texts	Syntactic parsing trees + linguistic rules	Rich syntactic-semantic analysis / language-specific
Odilinye et al. (2015)	Learning material	Sentence ranking + WordNet	Effective sentence selection / requires term definitions
Le and Pinkwart (2015)	Educational texts	Semantic similarity + lexical databases	Context-aware selection / limited lexical coverage
Shwe et al. (2024)	Learning content	Semantic similarity + ranking	Captures semantic relevance / complex pipeline
Das et al. (2021a)	Text corpus	RAKE + NER + K-means clustering	Generates MCQs / high computational cost, multi-step process

4.2.3 Transformer-Based Approaches

A deep learning approach that has gained prominence in research in recent years is Transformers. Transformers are a specific architecture of neural networks that learn context and, thus, meaning by monitoring relationships in sequential data, such as the words in a sentence, for example (Waswani et al., 2017). Among the studies focused on AQG analyzed in the last decade, 43 used Transformer-based approaches. The Transformer architecture is the driving force behind LLMs, which are AI models capable of understanding and generating natural language. These language models have been widely used for performing tasks typically humans, such as question generation Doughty et al. (2024); Matsumori et al. (2023)), for example.

Several strategies have been explored within this category:

- **Prompt Engineering (PE)**: LLMs are guided through prompts containing task instructions, learning objectives, or Bloom's Taxonomy categories (Doughty et al. (2024); Lee et al. (2023); Maity et al. (2024)). PE has been applied for MCQ generation and distractor design.
- **Information Extraction + Transformers**: Key sentences are extracted from educational materials and encoded into semantic vectors (e.g., via Sentence-BERT), which serve as input for question generation models such as T5 (H.-C. Wang et al. (2023); Jasuja et al. (2024)).
- **Masked Language Models (MLM)**: Fill-in-the-blank questions are generated by masking words and predicting candidates using BERT or similar models, often complemented with distractor filtering based on confidence scores (Andersson and Picazo-Sanchez (2023); Matsumori et al. (2023)).
- **Fine-tuning LLMs**: Pre-trained models are adapted to specific datasets, languages, or domains to improve task-specific performance (Awalurahman and Budi (2024b); Babakhani et al. (2024); Ruma et al. (2023)).

Transformer-based approaches demonstrate high linguistic flexibility, scalability, and the ability to produce questions comparable to human-generated ones (Bachiri and Mouncif (2023); Kyung-Mi (2024)). Limitations include dependency on computational resources, the need for domain-specific fine-tuning, and variable performance in low-resource languages (Ruma et al. (2023)). From an educational perspective, their capacity to generate high-quality and context-sensitive questions offers strong potential to support adaptive learning and reduce teachers' workload. Nevertheless, their complexity, opacity, and bias represent barriers to classroom adoption and may challenge educators' trust in automated assessments. Table 6 summarizes contributions, highlighting their input data, techniques, and limitations.

4.2.4 Ontology-Based Approaches

An approach that can be observed in studies published throughout the decade is ontologies. Ontologies are formal descriptions of a set of concepts within a domain and the relationships between these concepts (Jouault et al., 2016). In ontologies, the concepts within a specific domain are represented by classes. Classes are abstractions that group entities that share common characteristics.

Table 6: Representative Transformer-based approaches for automatic question generation.

Study	**Input / Data**	**Technique**	**Highlights / Limitations**
Doughty et al. (2024)	Programming logic texts	GPT-4 + Prompt Engineering	Comparable to human-generated MCQs
Matsumori et al. (2023)	Sentences	MLM (BERT)	Fill-in-the-blank / distractors filtered by confidence
H.-C. Wang et al. (2023)	Lecture notes / textbooks	Sentence-BERT + T5	Info extraction / semantic vector encoding
Rezigalla (2024)	PDFs (medical)	RAG + ChatPDF	Context-aware question generation / some domain-specific errors
Babakhani et al. (2024)	News articles	Fine-tuned T5 / GPT-3	Subjective question generation / dataset-specific training
Ruma et al. (2023)	Bengali QA datasets	Fine-tuned BanglaT5	Low-resource language adaptation / limited prior research

Based on this structure, the system developed by Jouault et al. (2016) uses the History Dependent Question Ontology (HDQ Ontology), which categorizes questions into different types, such as ”Description,” ”Method,” ”Explanation,” and ”Comparison.” Each question type is associated with specific classes. For example, a ”Description” question may be generated from a class that describes a historical event, while a ”Comparison” question may involve subclasses that compare different events or historical figures. The properties associated with the classes help contextualize the questions.

Alsubait et al. (2016) and Leo et al. (2019) used ontologies to generate MCQs in the biological and medical domains, respectively. Both approaches identify relevant concepts and generate a question for each, extracting the correct answer and appropriate distractors while controlling question difficulty. The plausibility of incorrect alternatives increases question challenge, and similarity measures between concepts (e.g., a Jaccard-inspired formula in Alsubait et al. (2016)) are used to select distractors. Cubric and Tosic (2020) extended this approach by combining ontologies with templates based on Bloom’s Taxonomy, highlighting that preprocessing improves language quality and user-friendly labeling. Kusuma et al. (2022) further classified ontologies into taxonomy, sentence, and knowledge types, merging them by identifying classes and instances with the same label.

To enrich ontologies with semantic information, Kumar et al. (2023) implemented rules using Description Logic and Semantic Rule Language. These rules establish hierarchical relationships and conditions that allow automatic inference of new knowledge, enabling the generation of MCQs based on concept interactions and properties. The combination of these methods improves the accuracy of inferences and ensures that generated questions are relevant, contextually appropriate, and effective for educational assessment. Table 7 summarizes these approaches, highlighting their domains, techniques, and main contributions.

Table 7: Representative ontology-based AQG approaches.

Study	**Domain / Data**	**Technique**	**Highlights / Limitations**
Jouault et al. (2016)	History	HDQ Ontology + question types	Structured questions: Description, Comparison, etc.
Alsubait et al. (2016)	Biology	Ontology + distractor similarity	Uses Jaccard-inspired formula to select plausible distractors
Leo et al. (2019)	Medicine	Ontology-based MCQs	Controlled difficulty via distractor selection
Cubric and Tosic (2020)	General / Bloom's Taxonomy	Ontology + templates	Aligns with cognitive levels / needs preprocessing
Kusuma et al. (2022)	General	Taxonomy, sentence, knowledge ontologies merged	Merges ontology types for richer question generation
Kumar et al. (2023)	Education	Ontology + semantic rules	Description Logic + Semantic Rule Language for context-aware MCQs

Overall, ontology-based approaches allow structured question generation and controlled distractor selection, but they require careful preprocessing of labels and integration with templates or rules to produce high-quality educational questions. From an educational perspective, these methods can support alignment with learning objectives, cognitive levels, and assessment standards, providing teachers with more precise and pedagogically relevant question sets. However, the need for domain-specific ontologies, preprocessing effort, and technical knowledge may limit their scalability and adoption in typical classroom settings.

4.2.5 Deep Learning-Based Approaches – Except Transformers

Some AQG approaches use other deep learning-based methods that are not Transformers. The work by Z. Wang et al. (2018) generated questions using a type of recurrent neural network (RNN) designed to address the issue that occurs in traditional RNNs when training on long sequences. The architecture used was LSTM. The main innovation of LSTMs is the introduction of memory cells, which allow the network to retain information for extended periods. Each memory cell can store information and decide when to keep or discard it. The model trained by Z. Wang et al. (2018) received as input contexts (e.g., a paragraph from an educational text) and an associated answer for each. The output is generated questions related to the content of the contexts. Liu et al. (2019) proposed an approach that combines copy mechanisms with neural networks to improve question generation by learning to identify relevant keywords in the text.

In the experiment proposed by Murad et al. (2023), the LSTM and GRU (Gated Recurrent Unit) approaches were compared. GRU is an RNN architecture that can be considered a simpler

and more efficient alternative to LSTM while maintaining the ability to capture long-term dependencies in sequential data. During the training process, the authors used a dataset consisting of 778 pairs of key phrases and 240 question templates, which were organized according to Bloom's taxonomy levels. The authors concluded that, although GRU may be more time-efficient, LSTM excelled in terms of the quality of the generated questions, making it more suitable for educational contexts that require a higher level of complexity and contextualization.

H.-C. Wang et al. (2024) also use the LSTM architecture in question generation, proposing a model called TE-QG (Teaching Evaluation Question Generation). The approach involves identifying words or terms that represent the main topics addressed in the text, using POS filtering to select relevant words. The experiments conducted by the authors demonstrated that the inclusion of these topic features significantly improves the model's performance, with the generated questions being rated as more fluent and relevant.

From an educational perspective, deep learning–based approaches, such as LSTM architectures, can generate context-aware and pedagogically relevant questions, potentially supporting teachers in assessment design and reducing manual workload. However, their reliance on annotated data, technical expertise, and computational resources may limit accessibility for some educational settings. These representative approaches are summarized in Table 8, which highlights the main architectures, techniques, and findings.

Table 8: Representative deep learning-based AQG approaches (excluding Transformers).

Study	**Architecture**	**Technique / Features**	**Highlights / Limitations**
Z. Wang et al. (2018)	LSTM	Context + answer input	Generates context-aware questions / memory retains long-term dependencies
Liu et al. (2019)	LSTM + copy mechanism	Keyword extraction	Improves question relevance by identifying important words
Murad et al. (2023)	LSTM / GRU	Bloom's taxonomy templates	LSTM better quality; GRU faster
H.-C. Wang et al. (2024)	LSTM (TE-QG)	POS-based topic features	Enhances fluency and relevance; suitable for educational assessment

4.2.6 Combined Architectures-Based Approaches and Comparative Analyses

In literature, we can observe a set of works that not only use one technique for question generation, but rather a combination of techniques in a combined approach. Furthermore, other works have aimed to conduct a comparative analysis between different techniques. These two approaches will be discussed in this subsection.

Khandait et al. (2023) propose a methodology that integrates sentiment analysis and structural analysis of input text, combined with preprocessing steps such as stemming and named entity recognition. Using DistilGPT2, the system generates diverse and meaningful questions, avoiding repetitions and improving accuracy. Alshboul and Baksa-Varga (2024) developed a framework for

generating open-ended questions about Python programming, using an approach that combines ontologies and Transformers: an ontology classifies elements of code snippets, enriching their semantic interpretation, before feeding the information into QuestGen AI for question generation. In a similar direction, Maheen et al. (2022) apply NLP preprocessing (tokenization, lemmatization, POS tagging) followed by BERT embeddings and clustering to identify informative sentences from textbooks, and then generate distractors using WordNet, Wiktionary, and Google search.

Other studies integrate graph-based techniques. Xu et al. (2023) information extraction with Graph Attention Networks (GAT) and GRUs, producing enriched contextual representations that improve question generation. Similarly, Z. Li et al. (2023) construct knowledge graphs from entities identified in text, integrating them with contextual representations via GAT before passing them to a multi-head attention model. Hsiao and Chung (2022) also follow this line, using RNNs and local knowledge graphs for semantic enrichment.

Seq2seq-based solutions are also common. Phan and Do (2022) adapt BERT for named entity recognition, converting entities into questions via seq2seq models. Chung et al. (2024) incorporate reinforcement learning to refine sequences based on quality metrics, while Yu et al. (2021) integrate contextual retrieval and transfer learning across languages to improve performance with limited data.

Hybrid approaches also mix traditional NLP with machine learning. Blšták and Rozinajová (2022) use a "composite pattern" structure to represent sentences before training a model to learn transformation rules. Comparative works further broaden this picture: Vincentio and Suhartono (2022) evaluate RNN-based (BiGRU, BiLSTM) and Transformer-based (mBART, mT5) models for Indonesian, finding clear superiority of Transformers, consistent with other studies (Xin et al. (2021)). Likewise, Sewunetie and Kovács (2022) show that template-based methods are outperformed by machine learning techniques such as multilayer perceptron (MLP). More recently, Suhartono et al. (2024) compare fine-tuned IndoBERT and IndoGPT for Indonesian AQG, concluding that IndoBERT achieves the best scores, surpassing even BiLSTM baselines.

From an educational perspective, combined AQG approaches offer the benefit of leveraging complementary strengths of different techniques, generating more diverse, context-aware, and pedagogically relevant questions. This can support adaptive learning, improve assessment quality, and reduce teachers' workload. At the same time, these approaches often require complex pipelines, technical expertise, and substantial computational resources, which may limit their adoption and scalability in typical classroom settings. These representative approaches and comparative analyses are summarized in Table 9, which highlights their techniques and key findings.

4.3 Answering RQ2 "What approaches are focused on BPM education?"

Computer Science stands out as one of the fields where AQG techniques have been most frequently proposed, ranking as the third most common, as shown in Table 3. Despite the relative popularity of studies in this area, no research has yet addressed question generation applied to BPM, with existing works remaining mostly focused on the teaching of programming languages. In the BPM education field, however, efforts have instead concentrated on other , including current and innovative teaching methodologies (Silva & Thom, 2021), as evidenced by contributions presented at the BPM Education Forum of the BPM 2024 conference (Marrella et al., 2024). These studies explore a wide range of strategies and experiences, including active methodologies such as

Table 9: Representative combined approaches and comparative analyses in AQG.

Study	**Techniques Combined / Compared**	**Highlights / Observations**
Khandait et al. (2023)	DistilGPT2 + sentiment + structural analysis	More diverse and meaningful questions
Alshboul and Baksa-Varga (2024)	Ontologies + Transformer	Open-ended Python questions; semantic enrichment
Maheen et al. (2022)	NLP preprocessing + BERT + clustering + WordNet	Extracts informative sentences; generates semantically related distractors
Xu et al. (2023)	GRU + GAT + knowledge graphs	Contextually enriched representations improve question generation
Phan and Do (2022)	BERT + seq2seq	Converts named entities into questions
Vincentio and Suhartono (2022)	BiGRU/BiLSTM vs mBART/mT5	Transformers outperform RNNs in Indonesian AQG
Sewunetie and Kovács (2022)	Template-based vs MLP	MLP produces higher-quality questions
Suhartono et al. (2024)	IndoBERT vs IndoGPT vs BiLSTM	Fine-tuned IndoBERT achieves best evaluation scores

Flipped Learning and Project-Based Learning, practical consulting projects, e-learning initiatives, eye-tracking methods, ERP teaching, design spaces for process redesign, and, more recently, the integration of large LLMs. Together, these contributions illustrate that BPM education has been an active field of pedagogical experimentation and innovation. AQG, however, has not yet been explored in this context.

Although Information and Communication Technologies (ICT) applied to education have advanced in recent years, the literature has previously pointed out the lack of educational initiatives focused on BPM that integrate such technologies. Studies such as those by Sarvepalli and Godin (2017), Wiechetek et al. (2017), Caporale et al. (2013), and Seethamraju (2012) have highlighted, among other gaps, the limited incorporation of IT systems in the BPM teaching-learning process, the absence of interactive resources such as games, and the lack of integrated and interdisciplinary curricula. Furthermore, Chow (2021) identified methodological shortcomings in the teaching of the subject, including a lack of student-centered approaches and active learning methodologies, which often rely on ICTs.

This scenario helps explain the scarcity of recent initiatives that integrate innovative approaches, such as AQG, in BPM education. The limitations already identified in the literature, combined with the absence of AQG-related studies in the past decade, not only highlight the historical challenges of the field but also suggest that these gaps still persist.

Based on the review of AQG applications in other fields, it is possible to envision how such

techniques could be applied to BPM education. In such a framework, key learning objectives and topics from BPM courses would serve as the foundation for automatically generating assessment items and practice questions. These questions could then be delivered through interactive platforms, supporting quizzes, simulations, or other active learning strategies. The framework would allow questions to be dynamically adapted to students' progress, providing timely feedback and personalized learning opportunities. In this way, AQG could address the lack of interactive and student-centered resources identified in the literature, while fostering engagement and reinforcing understanding of BPM concepts.

Although this framework remains conceptual and requires empirical validation, it outlines a concrete pathway for bridging the gap between advances in AQG and the specific needs of BPM education. A similar approach has been successfully applied in other domains. For instance, Kumar et al. (2023) developed a hybrid framework combining semantic-based and machine learning techniques to automatically generate multiple-choice questions for technical subjects.

4.4 Answering RQ3 "What limitations do existing approaches present?"

Over the past decade, automatic question generation has evolved significantly, driven by advancements in NLP techniques and machine learning. Consequently, the limitations of these approaches have also changed over time, reflecting emerging challenges as new methods have been proposed.

At the beginning of the studied decade, between 2015 and 2017, the main difficulties were related to the reliance on restrictive grammatical rules and template-based approaches, which often resulted in syntactically incorrect or poorly diversified questions (Alsubait et al. (2016); Odilinye et al. (2015)). Additionally, the lack of adequate automated metrics to assess the complexity and quality of questions was a recurring concern (Jouault et al. (2016); Mostow et al. (2016)). Some studies from that period also highlighted language-specific challenges, such as Chinese, which features more flexible structures and fewer prepositions, making it difficult to automatically generate coherent questions (Lui et al., 2015).

Starting in 2018, with the rise of deep learning models, new limitations emerged. Neural network-based models required large amounts of high-quality training data, which restricted their application in specific domains or low-resource languages (Yu et al. (2021); Keklik et al. (2019)). Model interpretability also became a significant issue, as deep neural networks often functioned as "black boxes," making it difficult to understand the criteria used to generate each question (Blšták & Rozinajová, 2022). Another issue that arose during this phase was the tendency of models to overfit the training data, reducing their ability to generalize to new contexts (H.-C. Wang et al., 2023).

In recent years, between 2022 and 2024, the most evident challenges have involved the quality of generated questions and automated evaluation. Recent studies highlight the difficulty of generating truly diverse questions with precise meaning (Goyal et al. (2023); Khandait et al. (2022)). The reliance on LLMs has also introduced challenges such as data bias and high processing costs (Alshboul and Baksa-Varga (2024); Rezigalla (2024)). Moreover, recent research points to limitations in the models' ability to handle questions requiring logical reasoning and complex arithmetic (Pham et al., 2024), as well as the varying complexities of pedagogical objectives (Lohr et al., 2024). Finally, there is a recognized need for greater standardization in datasets and evaluation metrics for generated questions (Sewunetie & Kovács, 2022).

4.5 Answering RQ4 "How has Generative AI influenced the field of AQG?"

Generative AI has gained prominence in the field of AQG, to the point that 43 original studies found in this review exclusively used techniques involving this approach in question generation. The main Generative AI used in text generation, and consequently in question generation, are LLMs. To address RQ4, it is important to investigate the reasons for choosing Generative AI-based approaches in the production of AQGs. This section aims to answer what led the authors to choose Generative AI as a technique for AQG and what positive aspects were observed regarding this technique compared to others.

With the popularization of LLMs, some studies have used Generative AI not because they consider these models more effective in generating questions, but with the aim of testing and evaluating their ability to perform this specific task ((Lohr et al., 2024); (de-Fitero-Dominguez et al., 2024); (Grévisse et al., 2024)). The hypothesis of Elkins et al. (2024) was that, given the success of LLMs in other tasks, these models could generate questions from contexts provided by teachers. According to the authors, the experiments conducted demonstrated the potential of LLMs in creating questions aligned with the learning objectives defined in Bloom's Taxonomy.

According to Lee et al. (2023), LLMs have a good ability to generate questions, but this can be enhanced through PE. Recurring issues in language models, such as hallucinations that cause questions to be generated from inaccurate content, and the low variety and adaptability of generated questions, can be addressed through PE. According to the authors, this allows the questions generated by LLMs to better meet the specific needs of the educational context.

Kumar et al. (2023) justify the adoption of LLMs in their AQG architecture due to their ability to encapsulate large amounts of knowledge and their proficiency in understanding language, thus improving the quality and relevance of the automatically generated multiple-choice questions. According to Kalpakchi and Boye (2024), the use of LLMs is advantageous compared to other AQG techniques because of their ability to generate various question formats and handle more complex sentence structures than other approaches. According to the authors, this makes them ideal for creating educational content that engages students in a manner like what humans produce (Kalpakchi & Boye, 2024).

According to Lohr et al. (2024), LLMs are capable of generating content of better semantic quality than other AQG approaches, especially with the adoption of RAG, which allows for more effective integration of information relevant to the specific context. This technique enables the created questions to be more aligned with educational objectives and students' needs. Furthermore, RAG helps mitigate common issues in traditional approaches, such as the superficiality of questions, promoting greater depth and relevance in the questions.

In summary, Generative AI has proven to be a powerful tool in the field of AQG, especially through LLMs, which offer an impressive ability to generate questions aligned with educational objectives. The reviewed studies highlight not only the effectiveness of this approach but also its potential for adaptation and customization using techniques such as PE and RAG. With these innovations, Generative AI has significantly contributed to improving the quality, relevance, and depth of generated questions, demonstrating its positive impact on the advancement of AQG.

4.6 Answering RQ5 "What metrics or criteria are used to evaluate the quality of generated questions?"

The evaluation of the quality of automatically generated questions is a fundamental step in the development of automatic question generation systems. Over the past decade, various studies have proposed different approaches to measure the effectiveness of these questions, using both automatic metrics and human evaluations, and in some cases, combining both. We now proceed to analyze the metrics used in different studies.

Several studies have exclusively adopted automatic metrics to assess the quality of generated questions. The most common metrics include BLEU, ROUGE, METEOR, and F1-score, which are widely used to measure textual similarity and linguistic quality. Studies such as those by H.-C. Wang et al. (2023), H.-C. Wang et al. (2024), Fahad et al. (2024), Goyal et al. (2023), Sewunetie and Kovács (2022), Kusuma et al. (2022) and Babakhani et al. (2024) have employed these metrics to quantify the accuracy of questions in relation to a reference set. Some research has also explored more sophisticated metrics, such as BERTScore, BLEURT, and CIDEr (Babakhani et al. (2024); Kalpakchi and Boye (2024)), which better capture semantic relationships between generated and reference questions. In addition to similarity-based metrics, some studies have used more specific criteria for automatic evaluation. For example, Rezigalla (2024) and Mostow et al. (2016) applied difficulty and discrimination metrics for questions, while Matsumori et al. (2023) analyzed the correct response rate.

Another common approach in the literature is qualitative evaluation conducted by experts or users, considering aspects such as clarity, relevance, grammaticality, and difficulty. Studies such as those by Doughty et al. (2024), Hsiao and Chung (2022), Andersson and Picazo-Sanchez (2023) and Jouault et al. (2016) adopted this strategy, collecting subjective feedback on the quality of the generated questions. Some works specifically analyzed the answerability of the questions (Chinkina et al., 2020) or their suitability to the learning level, as in the case of Andersson and Picazo-Sanchez (2023), who considered the JLPT levels. Other studies focused on comparing automatically generated questions with human-produced ones. Bachiri and Mouncif (2023), Le and Pinkwart (2015) and Susanti et al. (2017) investigated the degree of similarity between the two sets, while Alsubait et al. (2016) and Larranaga et al. (2022) analyzed the acceptance of the questions by experts, without detailing specific metrics.

Finally, several studies combine automatic and subjective approaches to provide a more comprehensive analysis of the quality of the generated questions. Works such as those by Gašpar et al. (2023), Xu et al. (2023), Xin et al. (2021) and Ruma et al. (2023) applied metrics like BLEU, METEOR, and ROUGE alongside human evaluations of grammaticality, relevance, and complexity. Additionally, some studies used statistical methods to measure the degree of agreement between human evaluators, such as Cohen's Kappa coefficient (Mulla and Gharpure (2023b); Murad et al. (2023); Wijanarko et al. (2021)). Another interesting strategy is the comparison between automatic and human evaluations to validate the effectiveness of computational metrics. For instance, Rodrigues et al. (2022) e Das et al. (2021a), analyzed correlations between subjective evaluations and traditional metrics, aiming to identify which automatic criteria best represent the human perception of question quality.

4.7 Answering RQ6 "What resistance do experts have regarding the automatic generation of content questions, especially among education professionals?"

An essential phase in software development, especially for tools aimed at assisting professionals, is validation with domain experts. This process ensures acceptance and trust in the solution while providing value to stakeholders (Pereira et al., 2024). When it comes to tools that support the learning process, this trust becomes even more crucial, as it directly influences users' willingness to adopt them in educational contexts. If teachers and students do not trust the tool's accuracy, usability, and effectiveness, its acceptance and impact on learning may be significantly reduced. RQ6 aims to answer how the acceptance of AQG tools by domain experts has been reported in studies published over the past decade.

As answered by RQ4, it was observed that several studies proposed evaluating questions automatically generated by teachers and instructors. However, the focus of these studies was on using the expertise of specialists to analyze the quality of the questions in terms of clarity, accuracy, and difficulty level, without investigating whether these professionals would hesitate to adopt such tools in the classroom. The closest to addressing this issue were occasional statements indicating a willingness to use them, as seen in the works of Chinkina et al. (2020), Larranaga et al. (2022) and Bachiri and Mouncif (2023), for example.

In this sense, the scarcity of studies exploring potential resistance to the adoption of AQG systems becomes evident. Investigating the factors that may influence this acceptance, such as the reliability of the generated questions, the ease of integration into teaching practices, and the level of control offered to teachers, could contribute to a better understanding of the challenges and opportunities in implementing these tools in the educational context.

4.8 Summary of Results

This subsection aims to synthesize the main findings of the conducted SLR, providing a clearer overview of the data analyzed and facilitating the reader's understanding. The summarized results offer a structured perspective on the key aspects explored in the selected studies.

Table 10 provides a comprehensive summary of the results from the SLR of AQG studies. It consolidates findings from 90 primary studies and 13 reviews, highlighting the most common approaches, evaluation methods used, frequent limitations, the influence of Generative AI, and gaps and opportunities for future research.

The analysis of the consolidated results shows a significant diversity of approaches and methods used in AQG studies, with a focus on techniques such as Transformers, Deep Learning, and combined approaches. The evaluation metrics vary between automatic and qualitative, suggesting an attempt to balance the objectivity of numerical assessments with the subjectivity of human analysis. However, limitations such as dependence on large data volumes, interpretability issues, and model bias are recurring challenges, highlighting the need for more robust solutions. Furthermore, the growing influence of generative AI, particularly through techniques like PE and RAG, has shown a positive impact on generating questions aligned with educational objectives, although important gaps remain, such as the lack of studies on educators' potential resistance to adopting these tools and the application of AQG in contexts like BPM.

Table 10: Summary of Results.

Category	Description
Number of Analyzed Articles	90 primary studies and 13 reviews
Main Approaches	Template-based and rule-based approaches, NLP, Transformers, Ontology, Deep Learning, and Combined Architectures.
Quality Assessment	Automatic metrics (BLEU, ROUGE, METEOR) and qualitative assessments (clarity, relevance, difficulty), with some combinations of both.
Common Limitations	Dependence on extensive data, model interpretability issues, overfitting, data bias, high computational cost, and lack of standardization in evaluation metrics.
Influence of Generative AI	Ability to generate questions aligned with educational objectives. Techniques such as PE and RAG have enhanced the adaptation and quality of generated questions.
Gaps and Opportunities	Lack of studies on educators' resistance to adopting AQG tools and lack of AQG applications in BPM.

The reviewed AQG approaches reveal educational benefits, challenges, and barriers. Overall, AQG can reduce professors' workload, diversify assessment, and foster active learning, provided that implementation strategies address technical limitations and support educators' trust and training.

5 Conclusion

The diversity of approaches found in the literature highlights the complexity of the task of AQG. Over the past decade, a wide range of techniques for AQG has emerged, ranging from simpler methods based on templates and rules to more advanced approaches, such as the use of large language models.

Regarding the evaluation of the quality of automatically generated questions, while automatic metrics offer efficiency and scalability, human evaluations provide more subjective insights into aspects such as clarity and relevance. Recent studies indicate a trend toward hybrid approaches that combine the best of both worlds to achieve a more robust and reliable evaluation.

In summary, the evolution of AQG has been marked by a transition from structural and grammatical limitations to challenges related to quality, evaluation, and model generalization. Advances in NLP techniques and the development of more robust models are promising, but there are still issues to be addressed to ensure truly effective question generation adaptable to different contexts. Notably, there is a lack of research on AQG specifically focused on BPM, highlighting

a significant gap for future studies to explore.

5.1 Threats to Validity

This SLR may present some limitations. Among them is the time restriction, which may exclude classic studies that contributed to the development of the topic. Another possible limitation is related to language, as relevant research published in languages other than English and Portuguese may not have been considered, although English is the most widely used language in the global academic community. Another aspect to be highlighted as a potential limitation is the subjective nature of the selection and analysis of the studies, which may introduce selection bias. Additionally, the review focused on a specific set of databases, and other databases may contain relevant studies that were not captured. In particular, databases such as ERIC and Brazilian repositories and conference proceedings (e.g., RBIE, SBIE, WIE, WEI) were not included. Additionally, the review focused on a specific set of databases routinely used by our research group, and other databases may contain relevant studies that were not captured. Finally, potential publication bias should also be considered, as studies reporting negative or inconclusive results may be underrepresented. These limitations do not invalidate the review's findings but emphasize the importance of interpreting the conclusions with caution, considering the possible gaps and biases inherent in the adopted process.

5.2 Future Research Directions

For future work, we intend to further explore the potential applications of the techniques identified in this review within the BPM field, assessing their feasibility and impact in different contexts. Additionally, we aim to develop more well-defined methodologies tailored to the specific needs of the area, considering the particularities of learning and applying knowledge in BPM. Future research could also investigate strategies to overcome practical challenges in AQG adoption, such as scalability, integration with existing BPM tools, and evaluation of learning outcomes. Furthermore, exploring interdisciplinary collaborations and extending the scope to diverse educational and organizational contexts may provide deeper insights and enhance the generalizability of the findings.

Acknowledgements

This study was financed by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

References

Al Faraby, S., Adiwijaya, A., & Romadhony, A. (2023). Review on Neural Question Generation for Education Purposes. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-023-00374-x [GS Search].

Al Shuraiqi, S., Aal Abdulsalam, A., Masters, K., Zidoum, H., & AlZaabi, A. (2024). Automatic Generation of Medical Case-Based Multiple-Choice Questions (MCQs): A Review of Methodologies, Applications, Evaluation, and Future Directions. *Big Data and Cognitive Computing*, *8*(10). https://doi.org/10.3390/bdcc8100139 [GS Search].

Alshboul, J., & Baksa-Varga, E. (2022). A Review of Automatic Question Generation in Teaching Programming. *International Journal of Advanced Computer Science and Applications*, *13*(10), 45–51. https://doi.org/10.14569/IJACSA.2022.0131006 [GS Search].

Alshboul, J., & Baksa-Varga, E. (2024). A Hybrid Approach for Automatic Question Generation from Program Codes. *International Journal of Advanced Computer Science and Applications*, *15*(1), 10–17. https://doi.org/10.14569/IJACSA.2024.0150102 [GS Search].

Alsubait, T., Parsia, B., & Sattler, U. (2016). Ontology-Based Multiple Choice Question Generation. *KI - Künstliche Intelligenz*, *30*(2), 183–188. https://doi.org/10.1007/s13218-015-0405-9 [GS Search].

Andersson, T., & Picazo-Sanchez, P. (2023). Closing the Gap: Automated Distractor Generation in Japanese Language Testing. *Education Sciences*, *13*(12). https://doi.org/10.3390/educsci13121203 [GS Search].

Awalurahman, H., & Budi, I. (2024a). Automatic distractor generation in multiple-choice questions: A systematic literature review. *PeerJ Computer Science*, *10*. https://doi.org/10.7717/peerj-cs.2441 [GS Search].

Awalurahman, H., & Budi, I. (2024b). Paragraph vs Sentence in Automatic Question Generation Fine-Tuning using Text-to-Text Transfer Transformer for Bahasa Indonesia - Proc. Int. Conf. Educ. Technol. ICET. *Proc. Int. Conf. Educ. Technol. ICET*, 155–161. https://doi.org/10.1109/ICET64717.2024.10778465 [GS Search].

Babakhani, P., Lommatzsch, A., Brodt, T., Sacker, D., Sivrikaya, F., & Albayrak, S. (2024). Opinerium: Subjective Question Generation Using Large Language Models. *IEEE Access*, *12*, 66085–66099. https://doi.org/10.1109/ACCESS.2024.3398553 [GS Search].

Bachiri, Y.-A., & Mouncif, H. (2023). Artificial Intelligence System in Aid of Pedagogical Engineering for Knowledge Assessment on MOOC Platforms: Open EdX and Moodle. *International Journal of Emerging Technologies in Learning*, *18*(5), 144–160. https://doi.org/10.3991/ijet.v18i05.36589 [GS Search].

Bhadouria, A., Gupta, P., Bindal, P., Madan, K., & Sonal, S. (2024). Automated Examination System using Machine Learning and Natural Language Processing - Proceedings of the 2024 Sixteenth International Conference on Contemporary Computing. *Proceedings of the 2024 Sixteenth International Conference on Contemporary Computing*, 752–761. https://doi.org/10.1145/3675888.3676144 [GS Search].

Blšták, M., & Rozinajová, V. (2022). Automatic question generation based on sentence structure analysis using machine learning approach. *Natural Language Engineering*, *28*(4), 487–517. https://doi.org/10.1017/S1351324921000139 [GS Search].

Caporale, T., Citak, M., Lehner, J., Schoknecht, A., & Ullrich, M. (2013). Social bpm lab–characterization of a collaborative approach for business process management education. *2013 IEEE 15th Conference on Business Informatics*, 367–373. https://doi.org/10.1109/CBI.2013.60 [GS Search].

Ch, D. R., & Saha, S. K. (2023). Generation of Multiple-Choice Questions From Textbook Contents of School-Level Subjects. *IEEE Transactions on Learning Technologies*, *16*(1), 40–52. https://doi.org/10.1109/TLT.2022.3224232 [GS Search].

Ch, D. R., & Saha, S. K. (2020). Automatic Multiple Choice Question Generation From Text: A Survey. *IEEE Transactions on Learning Technologies*, *13*(1), 14–25. https://doi.org/10.1109/TLT.2018.2889100 [GS Search].

Chinkina, M., Ruiz, S., & Meurers, D. (2020). Crowdsourcing evaluation of the quality of automatically generated questions for supporting computer-assisted language teaching. *ReCALL*, *32*(2), 145–161. https://doi.org/10.1017/S0958344019000193 [GS Search].

Chow, W. (2021). Teaching business process management with a flipped-classroom and problem-based learning approach with the use of apromore and other bpm software in graduate information systems courses. *2021 IEEE International Conference on Engineering, Technology & Education (TALE)*, 1–8. https://doi.org/10.1109/TALE52509.2021.9678885 [GS Search].

Chung, H.-L., Chan, Y.-H., & Fan, Y.-C. (2024). Handover QG: Question Generation by Decoder Fusion and Reinforcement Learning. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, *32*, 3644–3655. https://doi.org/10.1109/TASLP.2024.3426292 [GS Search].

Corporation for Digital Scholarship. (2024). Zotero [Software for managing research sources]. https://www.zotero.org/

Cubric, M., & Tosic, M. (2020). Design and evaluation of an ontology-based tool for generating multiple-choice questions. *Interactive Technology and Smart Education*, *17*(2), 109–131. https://doi.org/10.1108/ITSE-05-2019-0023 [GS Search].

da Silva, D.-P., Ives, de Moura, Arruda, S., et al. (2023). Compreendendo a inteligência artificial generativa na perspectiva da língua. *SciELO Preprints*. https://doi.org/https://doi.org/10.1590/SciELOPreprints.7077 [GS Search].

Das, B., & Majumder, M. (2017). Factual open cloze question generation for assessment of learner's knowledge. *International Journal of Educational Technology in Higher Education*, *14*(1), 24. https://doi.org/10.1186/s41239-017-0060-3 [GS Search].

Das, B., Majumder, M., Phadikar, S., & Sekh, A. A. (2021a). Multiple-choice question generation with auto-generated distractors for computer-assisted educational assessment. *Multimedia Tools and Applications*, *80*(21), 31907–31925. https://doi.org/10.1007/s11042-021-11222-2 [GS Search].

Das, B., Majumder, M., Phadikar, S., & Sekh, A. A. (2021b). Automatic question generation and answer assessment: A survey. *Research and Practice in Technology Enhanced Learning*, *16*(1), 5. https://doi.org/10.1186/s41039-021-00151-1 [GS Search].

de-Fitero-Dominguez, D., Garcia-Cabot, A., & Garcia-Lopez, E. (2024). Automated multiple-choice question generation in Spanish using neural language models. *Neural Computing and Applications*, *36*(29), 18223–18235. https://doi.org/10.1007/s00521-024-10076-7 [GS Search].

Doughty, J., Wan, Z., Bompelli, A., Qayum, J., Wang, T., Zhang, J., Zheng, Y., Doyle, A., Sridhar, P., Agarwal, A., Bogart, C., Keylor, E., Kultur, C., Savelka, J., & Sakr, M. (2024). A Comparative Study of AI-Generated (GPT-4) and Human-crafted MCQs in Programming Education, 114–123. https://doi.org/10.1145/3636243.3636256 [GS Search].

Dumas, M., La Rosa, M., Mendling, J., Reijers, H. A., et al. (2013). *Fundamentals of business process management* (Vol. 1). Springer. https://doi.org/https://doi.org/10.1007/978-3-662-56509-4 [GS Search].

Elkins, S., Kochmar, E., Cheung, J., & Serban, I. (2024). How Teachers Can Use Large Language Models and Bloom's Taxonomy to Create Educational Quizzes - Proc. AAAI Conf. Artif.

Intell. (Wooldridge M., Dy J., & Natarajan S., Eds.). *Proc. AAAI Conf. Artif. Intell.*, *38*, 23084–23091. https://doi.org/10.1609/aaai.v38i21.30353 [GS Search].

Fahad, A. R., Al Nahian, N., Islam, M. A., & Rahman, R. M. (2024). Answer Agnostic Question Generation in Bangla Language. *International Journal of Networked and Distributed Computing*. https://doi.org/10.1007/s44227-023-00018-5 [GS Search].

Figshare. (2025, October). *Systematic literature review articles on bpm* (dataset; Version Modified on 2025-10-11) [Dataset exported from Rayyan containing records that remained after the first two exclusion criteria (duplicate removal and non-article filtering), as well as notes and screening decisions recorded during the review. Available at: https://figshare.com/s/2a9ec15fa1db6e370ee2]. dataset. figshare.

García-Méndez, S., De Arriba-Pérez, F., & Somoza-López, M. D. C. (2024). A Review on the Use of Large Language Models as Virtual Tutors. *Science & Education*. https://doi.org/10.1007/s11191-024-00530-2 [GS Search].

Gašpar, A., Grubišić, A., & Šarić-Grgić, I. (2023). Evaluation of a rule-based approach to automatic factual question generation using syntactic and semantic analysis. *Language Resources and Evaluation*, *57*(4), 1431–1461. https://doi.org/10.1007/s10579-023-09672-1 [GS Search].

Gorgun, G., & Bulut, O. (2024). Exploring quality criteria and evaluation methods in automated question generation: A comprehensive survey. *Education and Information Technologies*, *29*(18), 24111–24142. https://doi.org/10.1007/s10639-024-12771-3 [GS Search].

Goyal, R., Kumar, P., & Singh, V. P. (2023). Automated Question and Answer Generation from Texts using Text-to-Text Transformers. *Arabian Journal for Science and Engineering*. https://doi.org/10.1007/s13369-023-07840-7 [GS Search].

Grévisse, C., Pavlou, M. A. S., & Schneider, J. G. (2024). Docimological Quality Analysis of LLM-Generated Multiple Choice Questions in Computer Science and Medicine. *SN Computer Science*, *5*(5), 636. https://doi.org/10.1007/s42979-024-02963-6 [GS Search].

Hsiao, I.-H., & Chung, C.-Y. (2022). AI-infused Semantic Model to Enrich and Expand Programming Question Generation. *Journal of Artificial Intelligence and Technology*, *2*(2), 47–54. https://doi.org/10.37965/jait.2022.0090 [GS Search].

Hwang, W.-Y., & Utami, I. Q. (2024). Using GPT and authentic contextual recognition to generate math word problems with difficulty levels. *Education and Information Technologies*, *29*(13), 1–29. https://doi.org/10.1007/s10639-024-12537-x [GS Search].

Izo, F., Leão, J., Pirovani, J., & Oliveira, E. (2022). Automatic generation of large-scale assessment questions. *Anais do XVIII Simpósio Brasileiro de Sistemas de Informação*. https://sol.sbc.org.br/index.php/sbsi/article/view/21348 [GS Search].

Jadhav, P., Deore, S., Aute, S., Gaikwad, R., Barphe, S., & Meshram, E. (2024). 'Questomatic': Automated Question Formulation System - Int. Conf. Sustain. Expert Syst., ICSES - Proc. *Int. Conf. Sustain. Expert Syst., ICSES - Proc.*, 577–581. https://doi.org/10.1109/ICSES63445.2024.10763124 [GS Search].

Jasuja, H., Negi, U., & Kaur, G. (2024). Automatic Question Generation from Youtube Lectures using Deep Learning - Int. Conf. Comput. Commun. Netw. Technol., ICCCNT. *Int. Conf. Comput. Commun. Netw. Technol., ICCCNT*. https://doi.org/10.1109/ICCCNT61001.2024.10726107 [GS Search].

Jouault, C., Seta, K., & Hayashi, Y. (2016). Content-Dependent Question Generation Using LOD for History Learning in Open Learning Space. *New Generation Computing*, *34*(4), 367–394. https://doi.org/10.1007/s00354-016-0404-x [GS Search].

Kalpakchi, D., & Boye, J. (2024). Quinductor: A multilingual data-driven method for generating reading-comprehension questions using Universal Dependencies. *Natural Language Engineering*, *30*(2), 217–255. https://doi.org/10.1017/S1351324923000037 [GS Search].

Keklik, O., Tuglular, T., & Tekir, S. (2019). Rule-based automatic question generation using semantic role labeling. *IEICE Transactions on Information and Systems*, (7), 1362–1373. https://doi.org/10.1587/transinf.2018EDP7199 [GS Search].

Khandait, K., Bhura, S., & Asole, S. (2022). AUTOMATIC QUESTION GENERATION THROUGH WORD VECTOR SYNCHRONIZATION USING LAMMA. *Indian Journal of Computer Science and Engineering*, *13*(4), 1083–1095. https://doi.org/10.21817/indjcse/2022/v13i4/221304046 [GS Search].

Khandait, K., Bhura, S., & Asole, S. (2023). Neural Approach to Automatic Subjective Question Generation System Using Multiple Filters for Supporting Correct WH-type Question. *International Journal of Intelligent Systems and Applications in Engineering*, *11*(11s), 60–72. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85171370463&partnerID=40&md5=db08333b9f2c713c11bea6896b6f116a Publisher: Ismail Saritas.

Kitchenham, B., & Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering version 2.3 [Accessed: 2025-08-31]. *Engineering*, *45*(4ve), 1051. https://legacyfileshare.elsevier.com/promis_misc/525444systematicreviewsguide.pdf. [GS Search].

Kumar, A., Nayak, A., K, M., & Ghosh, K. (2023). A Novel Framework for the Generation of Multiple Choice Question Stems Using Semantic and Machine-Learning Techniques. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-023-00333-6 [GS Search].

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. https://doi.org/10.1007/s40593-019-00186-y [GS Search].

Kusuma, S., Siahaan, D., & Fatichah, C. (2022). Automatic question generation with various difficulty levels based on knowledge ontology using a query template. *Knowledge-Based Systems*, *249*. https://doi.org/10.1016/j.knosys.2022.108906 [GS Search].

Kyung-Mi, O. (2024). A comparative study of ai-human-made and human-made test forms for a university tesol theory course. *Language Testing in Asia*, *14*(1), 19. [GS Search].

Lafkiar, S., & En Nahnahi, N. (2024). An end-to-end transformer-based model for Arabic question generation. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-024-19958-3 [GS Search].

Larranaga, M., Aldabe, I., Arruarte, A., Elorriaga, J., & Maritxalar, M. (2022). A Qualitative Case Study on the Validation of Automatically Generated Multiple-Choice Questions From Science Textbooks. *IEEE Transactions on Learning Technologies*, *15*(3), 338–349. https://doi.org/10.1109/TLT.2022.3171589 [GS Search].

Le, N.-T., & Pinkwart, N. (2015). Evaluation of a question generation approach using semantic web for supporting argumentation. *Research and Practice in Technology Enhanced Learning*, *10*(1), 3. https://doi.org/10.1007/s41039-015-0003-3 [GS Search].

Lee, U., Jung, H., Jeon, Y., Sohn, Y., Hwang, W., Moon, J., & Kim, H. (2023). Few-shot is enough: Exploring ChatGPT prompt engineering method for automatic question generation in english education. *Education and Information Technologies*. https://doi.org/10.1007/s10639-023-12249-8 [GS Search].

Leo, J., Kurdi, G., Matentzoglu, N., Parsia, B., Sattler, U., Forge, S., Donato, G., & Dowling, W. (2019). Ontology-Based Generation of Medical, Multi-term MCQs. *International Journal of Artificial Intelligence in Education*, *29*(2), 145–188. https://doi.org/10.1007/s40593-018-00172-w [GS Search].

Li, E., Su, J., Sheng, H., & Wai, L. (2020). AGenT Zero: Zero-shot Automatic Multiple-Choice Question Generation for Skill Assessments. *arXiv preprint arXiv:2012.01186*. [GS Search].

Li, Z., Cao, Z., Li, P., Zhong, Y., & Li, S. (2023). Multi-Hop Question Generation with Knowledge Graph-Enhanced Language Model. *Applied Sciences (Switzerland)*, *13*(9). https://doi.org/10.3390/app13095765 [GS Search].

Liu, B., Zhao, M., Niu, D., Lai, K., He, Y., Wei, H., & Xu, Y. (2019). Learning to Generate Questions by LearningWhat not to Generate, 1106–1118. https://doi.org/10.1145/3308558.3313737 [GS Search].

Lohr, D., Berges, M., Chugh, A., Kohlhase, M., & Müller, D. (2024). Leveraging Large Language Models to Generate Course-Specific Semantically Annotated Learning Objects. *Journal of Computer Assisted Learning*, *41*(1). https://doi.org/10.1111/jcal.13101 [GS Search].

Lui, A. K.-F., Ng, S.-C., & Fung, Y.-C. (2015). A parse tree based computation technique for generating comprehension style questions from chinese text. *Technology in Education. Transforming Educational Practices with Technology: First International Conference, ICTE 2014, Hong Kong, China, July 2-4, 2014. Revised Selected Papers*, 61–73. https://doi.org/https://doi.org/10.1007/978-3-662-46158-7 [GS Search].

Madri, V. R., & Meruva, S. (2023). A comprehensive review on MCQ generation from text. *Multimedia Tools and Applications*, *82*(25), 39415–39434. https://doi.org/10.1007/s11042-023-14768-5 [GS Search].

Maheen, F., Asif, M., Ahmad, H., Ahmad, S., Alturise, F., Asiry, O., & Ghadi, Y. (2022). Automatic computer science domain multiple-choice questions generation based on informative sentences. *PeerJ Computer Science*, *8*. https://doi.org/10.7717/PEERJ-CS.1010 [GS Search].

Maity, S., Deroy, A., & Sarkar, S. (2024). Investigating Large Language Models for Prompt-Based Open-Ended Question Generation in the Technical Domain. *SN Computer Science*, *5*(8), 1128. https://doi.org/10.1007/s42979-024-03464-2 [GS Search].

Marrella, A., Resinas, M., Jans, M., & Rosemann, M. (Eds.). (2024). *Business process management: Blockchain, robotic process automation, central and eastern european, educators and industry forum* (Vol. 527). Springer. https://doi.org/10.1007/978-3-031-70445-1

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*, 9835–9850. https://doi.org/10.1109/ACCESS.2023.3239005 [GS Search].

Mostow, J., Huang, Y.-T., Jang, H., Weinstein, A., Valeri, J., & Gates, D. (2016). Developing, evaluating, and refining an automatic generator of diagnostic multiple choice cloze questions to assess children's comprehension while reading. *Natural Language Engineering*, *23*(2), 245–294. https://doi.org/10.1017/S1351324916000024 [GS Search].

Mulla, N., & Gharpure, P. (2023a). Automatic question generation: A review of methodologies, datasets, evaluation metrics, and applications. *Progress in Artificial Intelligence*, *12*(1), 1–32. https://doi.org/10.1007/s13748-023-00295-9 [GS Search].

Mulla, N., & Gharpure, P. (2023b). Leveraging well-formedness and cognitive level classifiers for automatic question generation on Java technical passages using T5 transformer. *International Journal of Information Technology*, *15*(4), 1961–1973. https://doi.org/10.1007/s41870-023-01262-2 [GS Search].

Murad, D., Wijanarko, B., Leandros, R., & Murad, S. (2023). LONG SHORT-TERM MEMORY-BASED ENCODER-DECODER WITH ATTENTION MECHANISM MODEL FOR QUESTION GENERATION. *ICIC Express Letters*, *17*(9), 1067–1074. https://doi.org/10.24507/icicel.17.09.1067 [GS Search].

Narayana, B., Harsha, V., Prudhvi, K., Vardhan, B., & Sravika, A. (2024). Automated Question Paper Generation using Natural Language Processing - Int. Conf. Comput. Intell. Green Sustain. Technol., ICCIGST - Proc. *Int. Conf. Comput. Intell. Green Sustain. Technol., ICCIGST - Proc.* https://doi.org/10.1109/ICCIGST60741.2024.10717510 [GS Search].

Neto, A. S., Silva, S., & Júnior, R. S. (2024). Chatgpt como ferramenta de aprendizagem: Um estudo com estudantes do ensino médio no interior do rn. *Anais do XXXV Simpósio Brasileiro de Informática na Educação*, 1480–1489. https://doi.org/10.5753/sbie.2024.241816 [GS Search].

Odilinye, L., Popowich, F., Zhang, E., Nesbit, J., & Winne, P. H. (2015). Aligning automatically generated questions to instructor goals and learner behaviour. *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)*, 216–223. https://doi.org/10.1109/ICOSC.2015.7050809 [GS Search].

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan—a web and mobile app for systematic reviews. *Systematic Reviews*, *5*(1), 210. https://doi.org/10.1186/s13643-016-0384-4 [GS Search].

Pereira, F. R., Costa, H. A. X., & Parreira, P. A. (2024). A Comparative Study of Tools for Anomaly Detection in Software Requirements, 11–21. [GS Search].

Pham, P. V. L., Duc, A. V., Hoang, N. M., Do, X. L., & Luu, A. T. (2024). ChatGPT as a Math Questioner? Evaluating ChatGPT on Generating Pre-university Math Questions. *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*, 65–73. https://doi.org/10.1145/3605098.3636030

Phan, T. H. V., & Do, P. (2022). NER2QUES: Combining named entity recognition and sequence to sequence to automatically generating Vietnamese questions. *Neural Computing and Applications*, *34*(2), 1593–1612. https://doi.org/10.1007/s00521-021-06477-7 [GS Search].

Reis, I., Fischer, C. E., Pereira, R., Juliani, D., & Ulbricht, V. (2024). A pedagogia digital a partir do conectivismo e o uso da inteligência artificial na educação: Uma revisão integrativa. *Anais do XXXV Simpósio Brasileiro de Informática na Educação*, 1689–1700. https://doi.org/10.5753/sbie.2024.242378 [GS Search].

Rezigalla, A. (2024). AI in medical education: Uses of AI in construction type A MCQs. *BMC Medical Education*, *24*(1). https://doi.org/10.1186/s12909-024-05250-3 [GS Search].

Rockembach, G. R., & Thom, L. H. (2024). Investigating the use of intelligent tutors based on large language models: Automated generation of business process management questions using the revised bloom's taxonomy. *Simpósio Brasileiro de Informática na Educação*

(SBIE), 1587–1601. https://doi.org/https://doi.org/10.5753/sbie.2024.242199 [GS Search].

Rodrigues, H., Nyberg, E., & Coheur, L. (2022). Towards the benchmarking of question generation: Introducing the Monserrate corpus. *Language Resources and Evaluation*, *56*(2), 573–591. https://doi.org/10.1007/s10579-021-09545-5 [GS Search].

Ruiz-Calleja, A., Vega-Gorgojo, G., Bote-Lorenzo, M., Asensio-Pérez, J., Dimitriadis, Y., & Gómez-Sánchez, E. (2021). Supporting contextualized learning with linked open data. *Journal of Web Semantics*, *70*. https://doi.org/10.1016/j.websem.2021.100657 [GS Search].

Ruma, J., Mayeesha, T., & Rahman, R. (2023). Transformer based Answer-Aware Bengali Question Generation. *International Journal of Cognitive Computing in Engineering*, *4*, 314–326. https://doi.org/10.1016/j.ijcce.2023.09.003 [GS Search].

Santi, M., Manacero, A., Peronaglio, F. F., Lobato, R. S., Spolon, R., & Cavenaghi, M. A. (2022). Training transformers for question generation task in intelligent tutoring systems. *2022 17th Iberian Conference on Information Systems and Technologies (CISTI)*, 1–6. https://doi.org/http://dx.doi.org/10.23919/CISTI54924.2022.9820606 [GS Search].

Sarvepalli, A., & Godin, J. (2017). Business process management in the classroom. *Journal of Cases on Information Technology (JCIT)*, *19*(2), 17–28. [GS Search].

Seethamraju, R. (2012). Business process management: A missing link in business education. *Business Process Management Journal*, *18*(3), 532–547. https://doi.org/https://doi.org/10.1108/14637151211232696 [GS Search].

Sewunetie, W. T., & Kovács, L. (2022). Comparison of Automatic Question Generation Techniques. *2022 IEEE 22nd International Symposium on Computational Intelligence and Informatics and 8th IEEE International Conference on Recent Achievements in Mechatronics, Automation, Computer Science and Robotics (CINTI-MACRo)*, 000025–000030. https://doi.org/10.1109/CINTI-MACRo57952.2022.10029559 [GS Search].

Sewunetie, W., & Kovacs, L. (2024). Automatic question generation using extended dependency parsing. *Indonesian Journal of Electrical Engineering and Computer Science*, *33*(2), 1108–1115. https://doi.org/10.11591/ijeecs.v33.i2.pp1108-1115 [GS Search].

Sewunetie, W., & Kovács, L. (2022). Comparison of template-based and multilayer perceptron-based approach for automatic question generation system. *Indonesian Journal of Electrical Engineering and Computer Science*, *28*(3), 1738–1748. https://doi.org/10.11591/ijeecs.v28.i3.pp1738-1748 [GS Search].

Shirude, A., Totala, S., Nikhar, S., Attar, V., & Ramanand, J. (2015). Automated Question Generation tool for structured data. *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 1546–1551. https://doi.org/10.1109/ICACCI.2015.7275833 [GS Search].

Shwe, L. L., Matayong, S., & Witosurapot, S. (2024). The unified difficulty ranking mechanism for automatic multiple choice question generation in digital storytelling domain. *Education and Information Technologies*. https://doi.org/10.1007/s10639-024-12666-3 [GS Search].

Silva, D., & Thom, L. (2021). 3d environment approach to teaching and learning business process management concepts: A systematic literature review. *Anais do XXXII Simpósio Brasileiro de Informática na Educação*, 328–337. https://doi.org/10.5753/sbie.2021.218159 [GS Search].

Suhartono, D., Majiid, M. R. N., & Fredyan, R. (2024). Towards automatic question generation using pre-trained model in academic field for Bahasa Indonesia. *Education and Information Technologies*. https://doi.org/10.1007/s10639-024-12717-9 [GS Search].

Susanti, Y., Tokunaga, T., Nishikawa, H., & Obari, H. (2017). Evaluation of automatically generated English vocabulary questions. *Research and Practice in Technology Enhanced Learning*, *12*(1), 11. https://doi.org/10.1186/s41039-017-0051-y [GS Search].

Thüs, D., Malone, S., & Brünken, R. (2024). Exploring generative AI in higher education: A RAG system to enhance student engagement with scientific literature. *Frontiers in Psychology*, *15*. https://doi.org/10.3389/fpsyg.2024.1474892 [GS Search].

Tuhpatussania, S., Utami, E., Kusrini, K., & Yuana, K. (2024). Automatic Question-Answer Generation for Education on Indonesian Texts : A Review of Methodologies, Dataset and Evaluation Metrics - Proc. - Int. Conf. Technol. Innov. Its Appl., ICTIIA. *Proc. - Int. Conf. Technol. Innov. Its Appl., ICTIIA*. https://doi.org/10.1109/ICTIIA61827.2024.10761325 [GS Search].

Vincentio, K., & Suhartono, D. (2022). Automatic Question Generation using RNN-based and Pre-trained Transformer-based Models in Low Resource Indonesian Language. *Informatica (Slovenia)*, *46*(7), 103–118. https://doi.org/10.31449/inf.v46i7.4236 [GS Search].

Wang, H.-C., Chiang, Y.-H., & Chen, I.-F. (2024). A Method for Generating Course Test Questions Based on Natural Language Processing and Deep Learning. *Education and Information Technologies*, *29*(7), 8843–8865. https://doi.org/10.1007/s10639-023-12159-9 [GS Search].

Wang, H.-C., Maslim, M., & Kan, C.-H. (2023). A question–answer generation system for an asynchronous distance learning platform. *Education and Information Technologies*, *28*(9), 12059–12088. https://doi.org/10.1007/s10639-023-11675-y [GS Search].

Wang, Z., Lan, A. S., Nie, W., Waters, A. E., Grimaldi, P. J., & Baraniuk, R. G. (2018). QG-net: A data-driven question generation model for educational content. https://doi.org/10.1145/3231644.3231654 [GS Search].

Waswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. [GS Search].

Weske, M., et al. (2007). Concepts, languages, architectures. *Business Process Management*. https://doi.org/https://doi.org/10.1007/978-3-642-28616-2 [GS Search].

Wiechetek, Ł., Medrek, M., & Banaś, J. (2017). Business process management in higher education. the case of students of logistics. *Problemy Zarzadzania*, (4/2017 (71)), 146–164. [GS Search].

Wijanarko, B. D., Heryadi, Y., Toba, H., & Budiharto, W. (2021). Question generation model based on key-phrase, context-free grammar, and Bloom's taxonomy. *Education and Information Technologies*, *26*(2), 2207–2223. https://doi.org/10.1007/s10639-020-10356-4 [GS Search].

Xin, Y., Cao, L., Wang, X., He, X., & Feng, L. (2021). Generating instructive questions from multiple articles to guide reading in e-bibliotherapy. *Sensors*, *21*(9). https://doi.org/10.3390/s21093223 [GS Search].

Xu, J., Sun, Y., Gan, J., Zhou, M., & Wu, D. (2023). Leveraging Structured Information from a Passage to Generate Questions. *Tsinghua Science and Technology*, *28*(3), 464–474. https://doi.org/10.26599/TST.2022.9010034 [GS Search].

Yu, J., Wang, S., & Yin, J. (2021). Adaptive Cross-Lingual Question Generation with Minimal Resources. *Computer Journal*, *64*(7), 1056–1068. https://doi.org/10.1093/comjnl/bxab106 [GS Search].

Zeghouani, O., Ali, Z., van Dijkhuizen, W., Hong, J., & Clos, J. (2024). AI in the Classroom: Examining the Feasibility of AI-Generated Questions in Educational Settings - ACM Int. Conf. Proc. Ser. *ACM Int. Conf. Proc. Ser.* https://doi.org/10.1145/3686038.3686652 [GS Search].

Zhang, R., Guo, J., Chen, L., Fan, Y., & Cheng, X. (2021). A Review on Question Generation from Natural Language Text. *ACM Trans. Inf. Syst.*, *40*(1). https://doi.org/10.1145/3468889 [GS Search].

Zimerman, F., Duarte, F. H., Silva, P. H., & Fortes, R. (2024). Explorando chatgpt para criação automática de questões práticas de programação de computadores. *Anais do XXXV Simpósio Brasileiro de Informática na Educação*, 2353–2364. https://doi.org/10.5753/sbie.2024.242440 [GS Search].