big data and cognitive computing

MDPI

Review

Automatic Generation of Medical Case-Based Multiple-Choice Questions (MCQs): A Review of Methodologies, Applications, Evaluation, and Future Directions

Somaiya Al Shuraiqi [1,*], Abdulrahman Aal Abdulsalam [1], Ken Masters [2], Hamza Zidoum [1] and Adhari AlZaabi [3]

[1] Department of Computer Science, College of Science, Sultan Qaboos University, P.O. Box 243, Muscat 123, Oman; a.aalabdulsalam@squ.edu.om (A.A.A.)
[2] Medical Education and Informatics Department, College of Medicine and Health Sciences, Sultan Qaboos University, P.O. Box 243, Muscat 123, Oman
[3] Department of Human and Clinical Anatomy, College of Medicine & Health Sciences, Sultan Qaboos University, P.O. Box 243, Muscat 123, Oman
* Correspondence: somaiya@squ.edu.om

Abstract: This paper offers an in-depth review of the latest advancements in the automatic generation of medical case-based multiple-choice questions (MCQs). The automatic creation of educational materials, particularly MCQs, is pivotal in enhancing teaching effectiveness and student engagement in medical education. In this review, we explore various algorithms and techniques that have been developed for generating MCQs from medical case studies. Recent innovations in natural language processing (NLP) and machine learning (ML) for automatic language generation have garnered considerable attention. Our analysis evaluates and categorizes the leading approaches, highlighting their generation capabilities and practical applications. Additionally, this paper synthesizes the existing evidence, detailing the strengths, limitations, and gaps in current practices. By contributing to the broader conversation on how technology can support medical education, this review not only assesses the present state but also suggests future directions for improvement. We advocate for the development of more advanced and adaptable mechanisms to enhance the automatic generation of MCQs, thereby supporting more effective learning experiences in medical education.

Keywords: automatic question generation (AQG); case-based multiple-choice questions (MCQs); ontology; natural language processing (NLP); machine learning (ML); deep learning (DL); large language model (LLM)

check for updates

Citation: Al Shuraiqi, S.; Aal Abdulsalam, A.; Masters, K.; Zidoum, H.; AlZaabi, A. Automatic Generation of Medical Case-Based Multiple-Choice Questions (MCQs): A Review of Methodologies, Applications, Evaluation, and Future Directions. *Big Data Cogn. Comput.* **2024**, *8*, 139. https://doi.org/10.3390/bdcc8100139

Academic Editors: Zuchao Li and Min Peng

Received: 5 June 2024
Revised: 13 October 2024
Accepted: 14 October 2024
Published: 17 October 2024

1. Introduction

Multiple-choice questions (MCQs) play an important role in automating the assessment of domain knowledge of students in a variety of disciplines [1,2]. In addition, the flexible administration of MCQs helps optimize knowledge assessment interventions. The manual development or generation of MCQs is cumbersome since it requires a high level of expertise to construct distractors or wrong options, the key or correct option, and the stem/question. Different MCQs, including odd-one-out, fill-in-the-blank, and wh-type, aim to examine the cognitive skills of students. To avoid the challenges of manual MCQ generation, machine learning and semantics-based approaches aid in the formulation of heterogeneous MCQ stems via automatic question generation (AQG). However, MCQs produced via automation need to be investigated for grammatical accuracy by semantics-oriented procedures, which reutilize the original content to produce Cloze/fill in the blank questions. Research suggests that combining machine learning and semantic-based techniques can be used to create multiple-choice questions (MCQs) with minimal grammatical and context problems [1].

The automation of MCQs in the medical education domain requires high-level skills and expertise in the medical domains [3,4]. To ensure the quality of automatically generated questions, various psychometric evaluation methods can be employed to assess their effectiveness. These methods help in determining the ability of the questions to discriminate between different levels of student knowledge, thereby enhancing the learning process. The evaluation of generated MCQs should be followed by a thorough psychometric analysis to validate their quality and ensure that they meet educational standards.

Medical case-based MCQs, in education, stand out for their emphasis on thinking skills over basic memorization [5]. These questions present scenarios that require students to analyze and synthesize information in order to arrive at solutions [6]. Studies suggest that using case-based assessment methods leads to performance and deeper learning compared to methods that focus on rote memorization [7]. Case-based learning in education aims to blend biomedical knowledge with real-world applications through patient cases, encouraging students to provide detailed responses rather than simple regurgitation of facts [8]. Moreover, the utilization of Case Based Reasoning (CBR) in Question Answering Systems showcases the effectiveness of matching cases with previous ones for accurate diagnoses, highlighting the value of case-based approaches in medical evaluations [9].

Recent evidence emphasizes a set of rules to effectively collate a reference set followed by the case-based compilation of the MCQs [10]. In addition, the assessment of the similarity of the parse structures of various case-based MCQs via automation improves their alignment and restructuring that aims to evaluate the domain knowledge of medical students. Importantly, the quality assessment of the automation of MCQs is highly necessary to nullify the risk of conceptual errors. The other improvement processes include feedback generation, presentation improvement, formulation of the automated template, and question structure enrichment [2].

The primary objective of this study is to provide academics and practitioners with a thorough summary of the research conducted in the domain of automated medical case-based MCQ generation. The study has several notable contributions, which are outlined as follows:

1. Provide a concise summary of the format of medical case-based multiple-choice questions (MCQs).
2. Offer a comprehensive examination of approaches used for generating medical case-based MCQs automatically.
3. Present applications of medical case-based MCQ auto-generation.
4. Give insight into evaluation and validation of automatically generated medical case-based MCQs.
5. Provide a concise overview of potential improvements and future research directions.

The structure of this paper is as follows: We first represent the structure of medical case-based MCQs and their components. Following that, we present a comprehensive explanation of methodologies for generating MCQs based on medical cases, as well as their applications. Then, we provide a comprehensive overview of the evaluation and validation of automatically generated MCQs. We finally identify the various research future directions and potential areas for further investigation.

2. Methods

The purpose of this paper is to analyze and synthesize the existing research on case-based MCQ generation, with a specific focus on the medical domain. This review aims to identify gaps in the current knowledge, highlight the progress made, and suggest directions for future research. The primary inquiry driving this analysis is as follows: what are the approaches employed to produce medical case-based multiple-choice questions (MCQs)?

We performed an extensive literature search on various databases, such as PubMed, Scopus, and Web of Science to ensure comprehensive and thorough coverage of important works in the field. The search strategy was designed to incorporate a blend of pertinent terms and phrases related to our subject matter, employing Boolean operators (‘AND’, ‘OR’) to enhance and concentrate the search. The search strategy incorporated a combination of keywords and phrases as follows: 'Automatic MCQs generation' AND 'medical case-based' OR 'item auto-generation', tailored to capture the broad spectrum of research on automatic generation of MCQs in the medical field. The search was limited to articles published in English and focused on the most recent developments in the field.

3. Background and Context

The development of automated multiple-choice questions (MCQs) has transformed significantly over time, evolving from a tool focused primarily on testing basic knowledge recall to one that assesses more complex cognitive skills in medical education [6,11]. Initially, MCQs were introduced to efficiently evaluate large numbers of students, providing a standardized method for assessing factual knowledge across various subjects. Their primary purpose was to test learners' ability to remember and reproduce information, a necessary but limited form of assessment. As medical education advanced and its demands became more intricate, the role of MCQs expanded. They are no longer restricted to simple knowledge recall but are now used to test higher-order thinking skills that are critical for medical professionals.

Today, MCQs in medical education encompass the evaluation of complex cognitive abilities, such as clinical reasoning, decision-making, and problem-solving [4,6]. In both formative and summative assessments, they have become important tools. In formative assessments, MCQs offer immediate feedback, allowing learners to identify and improve areas of weakness. In summative assessments, they provide a comprehensive measure of a student's knowledge and critical thinking, ensuring readiness for professional practice. Importantly, case-based MCQs can simulate real-life medical scenarios, requiring students to apply their knowledge in diagnosing conditions and making treatment decisions. This shift from assessing simple recall to evaluating critical thinking and decision-making has solidified MCQs as a crucial component of medical training, preparing students to navigate the complexities of healthcare.

Automating the generation of MCQs offers significant advantages over manual methods, primarily through its ability to quickly produce large volumes of questions, reducing the time and effort required by educators [12,13]. This automated approach not only saves time but also minimizes the potential for human errors in question creation. Automated MCQs can be tailored to align with specific educational objectives, ensuring that each question targets the intended learning outcomes and covers a broad range of cognitive skills, from basic recall to higher-order thinking [14]. This customization allows for more consistent and objective assessments, while also adjusting difficulty levels as needed. Additionally, automated MCQs are designed to present clear and concise questions, minimizing cognitive load on learners and allowing them to focus on the content being assessed, improving both the efficiency of the testing process and the overall learning experience.

Moreover, automated MCQ generation supports a variety of learning environments, including large-scale e-learning platforms and personalized learning programs [15]. In such settings, automatically generated MCQs can help learners assess their knowledge independently, facilitating continuous learning and self-assessment. This method of question generation is particularly beneficial in medical education, where learners must master a broad and constantly evolving body of knowledge.

Recent advancements in artificial intelligence (AI) and natural language processing have enabled the creation of automated MCQs that can assess higher-order thinking skills, such as critical thinking and problem-solving [1,16]. These AI-driven systems can analyze huge amounts of medical literature to generate questions that test a learner's ability to apply knowledge in real-world scenarios. Automated MCQs can include changing levels of difficulty and integrate distractors (incorrect answer choices) that challenge the learner's understanding of the material. This capacity to generate complex and nuanced questions makes automated MCQs a powerful tool in modern medical education.

However, automating MCQ generation presents several challenges, particularly in the medical field where the accuracy and relevance of questions are crucial [17]. The quality of automated MCQs largely depends on the system's ability to extract and structure relevant information from medical literature, generate plausible distractors, and stay updated with the latest research and clinical guidelines. These challenges are compounded by the complexity of medical knowledge and the rapid pace of advancements in the field. Another significant challenge is maintaining the semantic and structural consistency of questions. Automated systems must ensure that the MCQs they generate are aligned with the educational objectives and standards of the curriculum. In some cases, the complexity of medical knowledge can interrupt the system's ability to accurately capture and assess key concepts. Ensuring that automatically generated questions are free from bias and capable of evaluating a wide range of cognitive skills remains an ongoing area of research.

A study comparing the quality of MCQs generated through automated item generation (AIG) with those created manually found no significant differences in the overall quality or the cognitive domain assessed by the questions [16]. Both methods were able to produce high-quality questions that evaluated higher-order cognitive skills, demonstrating that automated systems can effectively complement traditional question generation methods. However, this study also highlighted the importance of careful oversight and validation in the automated generation process to ensure that the questions are both accurate and relevant.

In conclusion, while automated MCQ generation offers significant advantages in terms of efficiency and scalability, it also presents challenges that need to be addressed to ensure the quality and accuracy of the questions. As medical education continues to evolve, the integration of automated MCQs—especially those driven by AI—will likely play an increasingly important role in assessing both knowledge and critical thinking skills. However, ongoing research and refinement of these systems are crucial to overcoming the current limitations and ensuring that automated MCQs meet the high standards required in medical education.

4. Case-Based Reasoning (CBR) in Medical Education

Case-based reasoning (CBR) is an approach to problem-solving that uses historical cases to understand and address new problems [9,18,19]. In medical education, CBR is particularly valuable because it mirrors the real-life process of clinical decision-making. Physicians often draw on their knowledge of previous patient cases to diagnose and treat new patients with similar symptoms or conditions.

CBR involves several key steps:

1. **Understanding the Patient's Problem**: The first step in CBR involves thoroughly understanding the patient's symptoms and medical history, which helps in forming an initial idea about the possible medical conditions the patient might have.
2. **Knowledge Application**: Students apply their knowledge of anatomy, organ systems, and pathology to reason about the disease processes that could explain the patient's symptoms, which is crucial for accurate diagnosis.
3. **Pattern Recognition**: Students learn to recognize patterns in patient problems and compare them with illness scripts, which are mental representations of diseases based on previous cases they have studied or encountered.
4. **Systematic Discussion**: Through systematic discussion, students elaborate on the possible courses of action from the initial presentation of the patient to the final steps of clinical management, which helps in refining their clinical reasoning skills.
5. **Decision-Making Practice**: CBR also involves training students in decision-making from different perspectives, such as considering the burden on the patient and the cost for the hospital, which is essential for holistic patient care.
6. **Case Vignettes**: Students work with case vignettes that present different medical scenarios, helping them practice and apply their clinical reasoning skills in a controlled, educational environment.

This approach is beneficial in medical education, as it helps students develop critical thinking and diagnostic skills by applying theoretical knowledge to practical, real-world scenarios. CBR encourages deeper understanding and retention of medical knowledge by contextualizing learning within realistic clinical cases.

5. Structure of Case-Based MCQs

Case-based MCQs are a popular type of MCQ used in medical education and licensing exams to assess medical graduates' skills. A study of 1750 questions used in the German National Medical Licensing Exam found that 51.1% were case-based questions [20]. These questions are classified as testing higher-order thinking and problem-solving, discriminating better between low- and high-information students, and teaching pattern recognition skills. They have also been used to measure health professionals' adherence to clinical practice guidelines and approximate costly approaches for measuring clinical decisions. However, computerized generation of case-based questions is challenging due to their structured format and semantically incoherent stems.

Case-based MCQs follow a standard format, consisting of the stem (question), alternatives (options), and answer (correct answer) [4]. The stem should be well-defined and focused, containing the primary concept. Alternatives include all items from which the user must select one. The answer, also known as the 'correct answer' or 'key', must be indisputable and validated by referencing a reliable source for quality control purposes. However, this citation is used during the question development process and is not included in the item presented to examinees. Ambiguous phrases such as 'frequently', 'often', 'rarely', or 'sometimes' should be avoided in MCQs, as they can lead to confusion and misinterpretation. These phrases do not contribute to assessing the student's knowledge accurately and may instead test the student's ability to navigate ambiguities rather than their understanding of the content. The focus should be on crafting clear and precise questions that unambiguously assess the student's knowledge and skills. Distractors, which are all alternatives that are not the correct answer, are common in health sciences testing, with three or four being more common. Writing plausible distractors can be challenging when developing a well-formulated examination. Table 1 provides a real case-based MCQ from the National Board of Medical Examiners (NBME 2021) displaying the components of a case-based MCQ [21].

Table 1. The constituents of a case-based MCQ. Note that the answer (key) and the distractors together form the alternatives.

Case-Based MCQ Example	Constituent
"A 23-year-old man comes to the physician because of a 1-week history of painful urination and a clear urethral discharge. One month ago, he had similar symptoms and completed a course of doxycycline therapy for a chlamydial infection. He has no previous history of sexually transmitted diseases. He has been sexually active with one female partner for 2 years, and she takes an oral contraceptive. The examination shows no abnormalities. A urine polymerase chain reaction test is positive for Chlamydia trachomatis. Which of the following is the most likely explanation for this patient's current infection?" [21]	**Stem**
A. Concurrent infection with *Neisseria gonorrhoeae*	**Distractor 1**
B. Doxycycline-resistant strain of *C. trachomatis*	**Distractor 2**
C. Insufficient duration of therapy	**Distractor 3**
D. Reacquisition of infection from his partner	**Answer (or key)**
E. Sequestration of *C. trachomatis* in the epididymis	**Distractor 4**

6. Approaches for MCQ Generation

Creating manual MCQs requires estimation by experts in specific fields, which is a time-consuming process. Therefore, use of automatic approaches for producing MCQs from text is a viable alternative. Automatic generation is based on useful units using a diversity of methods.

6.1. Introduction to Medical Ontologies

Medical ontologies are structured frameworks that organize information in the medical field [6]. They offer a systematic approach to managing medical knowledge by establishing relationships between concepts such as diseases, symptoms, treatments, and patient data. This organization is essential for various applications, including clinical decision support, medical research, and the generation of educational content.

Medical ontologies can be classified into three main categories [22–24]:

1. **Field-Specific Ontologies:** Focused on particular areas of medicine, such as gene ontology (GO) and human phenotype ontology (HPO), these ontologies explore specific topics like gene functions and phenotypic abnormalities.
2. **General Medical Knowledge Ontologies:** These include comprehensive terminologies like Systematized Nomenclature of Medicine—Clinical Terms (SNOMED CT) and International Classification of Diseases (ICD), which cover a broad range of diseases, clinical findings, and procedures.
3. **Ontologies Addressing Common Misconceptions:** These are designed to clarify and correct frequently misunderstood information in medical fields, thus preventing errors in clinical practice and improving patient outcomes.

Overall, medical ontologies play a crucial role in providing a structured representation of medical knowledge. They support clinical decision-making, facilitate research, and contribute to the development of accurate and reliable educational materials.

6.2. Ontology-Based Approaches

Within the field of medical education, the creation of MCQs plays a crucial role in assessing and strengthening knowledge. Ontology-based systems have been a fundamental approach in automatically generating MCQs among many approaches used [25]. These methods are important in establishing questions that measure a range of medical information and are based on pre-established templates and rules.

Ontology-driven systems utilize sets of guidelines and patterns to analyze data sourced from a variety of references, like textbooks, academic papers or medical protocols [26]. MCQs are crafted in accordance with standards encompassing the question itself, the response, and multiple plausible but incorrect choices. These components play a role in the sector, where queries often mirror real-world scenarios or educational objectives crucial for training healthcare professionals.

One key advantage of using ontology-based systems is their ability to consistently generate high-quality questions [6]. This consistency ensures that the questions maintain a high standard, which is crucial for MCQs, as it helps ensure the validity and reliability of assessments in gauging students' understanding. Moreover, ontology-based systems can be developed without the need for advanced AI technologies or complex computational models. By relying on structured medical knowledge and logical relationships, these systems can produce high-quality MCQs that are both accessible and implementable, even in environments with limited AI infrastructure.

However, using ontology-based systems to create MCQs also presents several challenges. A significant issue is the reliance on the input criteria, which must be of high quality and depth. This necessitates the involvement of field experts to ensure that the generated questions are both relevant and educationally meaningful [27]. Subject matter experts play a crucial role in ensuring that the questions are not only medically accurate but also pedagogically sound. Additionally, regulatory requirements and predefined formats can limit the diversity and complexity of the MCQs. For instance, certain standards might require specific question structures, such as single-best-answer formats, which can restrict the ability to create questions that assess higher-order cognitive skills like reasoning and critical thinking [6].

Additionally, while these systems are effective at generating information-based questions, they may struggle with crafting questions that assess higher-order skills such as thinking and problem-solving [3]. For instance, an automated system might generate a basic factual question like 'what is the causative agent of measles?', which tests recall knowledge about measles disease. However, generating a more complex clinical scenario, such as 'A 5-year-old patient presents with a 3-day history of fever followed by a rash that started on the face and spread downwards. What is the most likely diagnosis?', requires a deeper understanding of clinical reasoning and decision-making. In this case, the options might include diagnoses like the common cold, measles, chickenpox, or rosacea, with 'measles' being the correct answer. Such questions demand the integration of medical knowledge and symptom analysis, which automated systems may struggle to produce without proper contextual understanding. This highlights the importance of utilizing a method that combines ontology-based systems with strategies to effectively address the complexities of decision-making when creating questions.

The combination of ontology-based systems, natural language processing (NLP), and machine learning (ML) helps ensure that the automated generation of MCQs follows a structured and consistent framework [2,27]. Ontology-based systems provide a systematic approach by defining the relationships between medical concepts, ensuring that the generated questions remain contextually accurate and relevant to the domain. Meanwhile, NLP and ML allow for the dynamic generation of questions that maintain uniformity in terms of complexity and relevance across different assessments. By leveraging these technologies, educators can ensure that the MCQs are not only consistent in terms of quality and difficulty but also align with specific educational goals, reducing human biases and variability in question creation. This systematic approach improves the reliability and fairness of the assessment process.

6.2.1. How Does Ontology Work to Generate MCQs?

Ontology is a detailed description of a domain, used to build intelligent applications and educational tools. It can be created using software like Protégé or languages like OWL (Web Ontology Language) [1]. There are many components of ontology design that should be considered to generate MCQs:

- **Ontology Components:**

Ontology includes concepts (classes), instances (individuals), attributes, relations, and axioms. Concepts are the abstract groups or categories that describe entities within a domain [28]. In the generation of medical case-based MCQs, classes might include 'Patient', 'Symptoms', 'Diseases', 'Treatments', and 'Tests'. These classes represent a group of entities sharing common characteristics. Also, instances are the specific entities or examples of classes. For example, a patient (A) or a disease (measles) are instances of the respective classes 'Patient' and 'Diseases'. In addition, attributes are the properties or features that describe the characteristics of concepts. For example, a patient might have attributes like age, gender, and medical history. Moreover, relations define how concepts are related to each other. In the medical ontology, relations can describe interactions like 'hasSymptom', 'hasTestResult', and 'givenTreatment'. These relationships help in linking different concepts together, providing a holistic view of the domain. Axioms are the rules that define the properties of concepts and their relationships. They help in asserting the truth about the concepts within the ontology, ensuring consistency and logical coherence. Figure 1 shows an example of medical ontology on COVID-19 [28]. It represents the relationships and hierarchical structure of concepts related to COVID-19. It illustrates how different entities such as viruses, symptoms, treatments, and medical departments are interconnected.

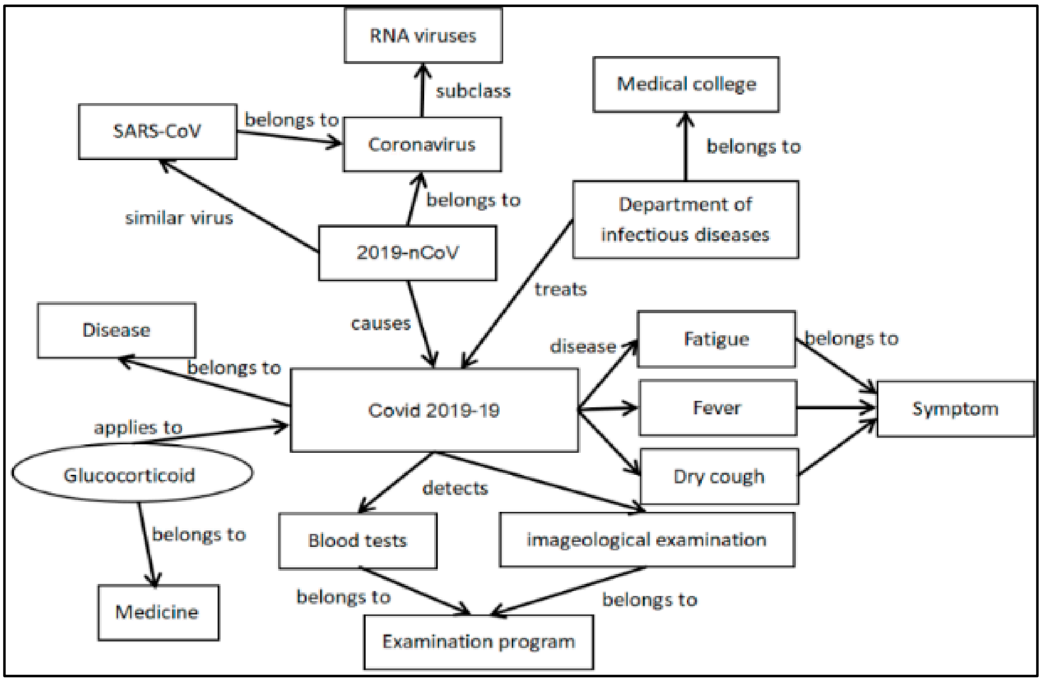

Figure 1. An example of medical ontology on COVID-19.

6.2.2. Rule-Based Generation

The rule-based generation technique combines popular ontologies used to represent domain information with rules that capture implicit knowledge or reveal new knowledge [29,30]. The rules within the ontology can use logical reasoning to generate new facts or initiate specified actions. Table 2 illustrates how ontology components, such as entities, properties, and rules, can be used to create scenarios in the medical domain, which can then be utilized for generating MCQs based on a rule-based approach.

Table 2. Designing a rule-based ontology in the medical domain for generating MCQs.

Ontology Design	Example in Medical Domain
Entities	Patient, Symptoms, Diseases, Treatments, Tests, etc.
Properties	'hasSymptom', 'hasTestResult', 'givenTreatment', etc.
Rules	These are the critical aspects of the rule-based ontology. For instance, a rule could be: "If a patient hasSymptom 'X' and hasTestResult 'Y', then they might haveDisease 'Z'."

- **Ontology Design**: The above table shows how to design an ontology using a rule-based method for creating MCQs based on medical cases [1]. Ontology design involves structuring and defining entities, properties, and rules to represent knowledge in the medical domain. Entities are the core components or concepts within the domain that are relevant to the creation of MCQs. In the medical domain, examples of these entities include patients, who represent the individuals receiving medical care; symptoms, which are signs or indications of a condition or disease; diseases, which are medical conditions that affect the patient; treatments, which are medical interventions used to treat diseases; and tests, which are diagnostic procedures used to identify or monitor diseases.
- **Properties**: Properties describe the attributes or relationships between entities. In the medical domain, examples include the property hasSymptom, which indicates the symptoms experienced by the patient; hasTestResult, which denotes the results of

diagnostic tests conducted on the patient; and givenTreatment, which specifies the treatments administered to the patient.

- **Rules**: Rules are guidelines that explain how entities and properties work together to create new knowledge or actions. They play a crucial role in a rule-based ontology system. For example, in the field of medical domain, a rule could be as follows: If a patient shows symptoms 'X' and tests positive for 'Y', then they could potentially have condition 'Z'. This rule utilizes information about symptoms and test results to suggest an illness based on the patient's situation. For instance, if a patient has a temperature and tests positive for a specific virus, this rule could indicate that the patient might have an illness linked to those symptoms and test outcomes.

Table 3 presents an example of the generation of medical MCQs based on cases using a rule-based ontology. An ontology rule is outlined, stating that "If a patient hasSymptom 'fever' and 'rash', they might haveDisease 'measles'." This rule is applied to craft an MCQ, where a 5-year-old patient presents with a 3-day history of fever followed by a rash that started on the face and spread downwards. The generated MCQ poses the question (stem) "What is the most likely diagnosis?" with answer choices including (A) Common cold, (B) Measles (correct), (C) Chickenpox, and (D) Rosacea. This method exemplifies how precise rules can be leveraged to produce educational MCQs based on clinical scenarios, ensuring the questions are relevant and medically accurate.

Table 3. Example of medical case-based MCQ generation by rule-based approach.

Ontology Rule:	"If a patient hasSymptom 'fever' and 'rash', they might haveDisease 'measles'."
MCQ Generation:	**Scenario:** "A 5-year-old patient presents with a 3-day history of fever followed by a rash that started on the face and spread downwards."
	Stem (Question): "What is the most likely diagnosis?"
	Options: (A) Common cold (B) Measles (correct) (C) Chickenpox (D) Rosacea

6.2.3. Template-Based Generation

The process of creating MCQs by utilizing pre-established templates is referred to as the template-based generation approach [6,22]. This method involves creating templates with structures allowing for the replacement of elements with appropriate content. One way to create a template is to have a question format that includes spaces where you can insert relevant phrases or ideas pertaining to the subject matter. This method simplifies the process of forming questions by incorporating information into existing templates. While it guarantees uniformity in question structure, it might lack the adaptability needed for addressing difficult issues. In Table 4, you can see how ontology architecture is developed for generating multiple-choice questions based on cases using a template-oriented approach. Table 5 demonstrates how MCQs are produced through a template-centric method.

Table 4. Creating a structure, for MCQ generation based on templates, in ontology design.

Ontology Design	**Example in Medical Domain**
Entities	Diseases, symptoms, treatments, etc.
Attributes	Characteristics of the diseases like onset time, severity, etc.
Relations	Links between entities, like a disease causing certain symptoms.

Table 5. A sample of creating MCQs based on medical cases using a template-driven method.

MCQ Template:	"A patient presents with [Symptom1], [Symptom2], and [Symptom3]. What is the most likely diagnosis?"
Entities:	Diseases, e.g., common cold, influenza, allergies
Attributes:	**Common Cold**: sneezing, runny nose, mild fever. **Influenza:** high fever, muscle aches, fatigue. **Allergies:** sneezing, itchy eyes, runny nose.
MCQ Generation:	"A patient presents with sneezing, runny nose, and mild fever. What is the most likely diagnosis?"
	Correct answer: Common Cold
	Distractors (wrong choices): Influenza, Allergies

Generating MCQs using both template- and rule-based approaches requires NLP techniques to create MCQs from text content [26,27,31–33]. NLP typically entails analyzing and understanding the material, pinpointing concepts or details, and then formulating questions based on this knowledge. NLP algorithms can assess sentences, recognize subjects and objects, and leverage these data to craft questions and distractors. This approach is particularly useful for crafting MCQs that closely align with the content of the text, ensuring that the questions are contextually relevant and accurate. An example of generating case-based MCQs using an NLP-based approach is illustrated in Table 6.

Table 6. Example of medical case-based MCQ generation by NLP-based approach.

Case Report	"A 45-year-old male patient presents with fever, cough, and difficulty breathing. Chest X-ray revealed bilateral pneumonia. The patient traveled recently to a region with a high number of COVID-19 cases".
Data Preprocessing	• **Tokenization:** ["A", "45-year-old", "male", "patient", "presents", "with", "fever"...]
Information Extraction	• Age: 45 • Gender: Male • Symptoms: Fever, cough, difficulty breathing • Findings: Bilateral pneumonia • History: Traveled to a high-risk COVID-19 region
MCQ Template Creation:	"A [age] year old [gender] with a history of [history] presents with [symptom]. What is the most likely diagnosis?"
Stem Generation:	"A 45-year-old male with a history of traveling to a high-risk COVID-19 region presents with fever, cough, and difficulty breathing. What is the most likely diagnosis?"
Distractor Generation:	• **Correct answer:** COVID-19 • **Distractors** (using NLP or medical knowledge base): Influenza, Tuberculosis, Common Cold
Generated MCQ:	"A 45-year-old male with a history of traveling to a high-risk COVID-19 region presents with fever, cough, and difficulty breathing. What is the most likely diagnosis?" A. Influenza B. Tuberculosis C. Common Cold D. COVID-19

6.2.4. Case-Based MCQ Based on Ontology Applications

Ontology-based systems for generating case-based MCQs represent a significant advancement in medical education. By leveraging structured information in ontologies, these systems create questions that closely mimic real clinical situations, enhancing the learning experience [4,6,11]. These systems provide detailed representations of medical cases, which help students and professionals assess their diagnostic and clinical decision-making skills more effectively. This method is particularly useful for equipping medical students with the necessary knowledge and skills for real-world patient interactions and for supporting the continuous professional development of healthcare practitioners, ensuring their expertise remains current with the latest advancements in medical procedures. Table 7 displays five distinct studies focused on MCQ-generation approaches in the medical field, including information on the techniques employed and the datasets linked with each study.

Table 7. Different studies in medical case-based MCQ auto-generation.

Study Reference	Techniques Used	Performance	Dataset Description
Karamanis et al. (2006) [34]	Rule-based approaches: • The pilot study uses a semi-automatic approach to generate multiple-choice test items (MCTIs) from medical text, based on Mitkov et al.'s work [35]. The system detects important concepts automatically and generates MCTIs testing factual knowledge. It uses the tf.idf method to promote key terms, computes distractors, and selects the best scoring distractors.	The average time taken per MCTI was around 3 min, which is significantly faster than manual production estimates by experts.	Text and the Unified Medical Language System (UMLS)
Wang et al. (2008) [26]	Template-based approaches: • The paper presents an automatic question generation system using medical articles and MMTx tools. The system extracts medical terms and classifies them using Unified Medical Language System (UMLS) and MetaMap Transfer (MMTx). It then selects similar templates and generates questions and answers. Participants can determine the interestingness of the questions, allowing dynamic question generation for online assessment and immediate feedback in medical learning systems.	Experiments conducted on 100 medical articles using 23 question templates on headache aspects showed 88 accurate questions generated, with 83 correctly answered. Mistakes in question generation were mainly due to insufficiently defined entries and keywords in templates.	A mix of text in diseases, symptoms, causes, therapies, medicines and devices.
Gierl et al. (2012) [36]	Template-based approaches: • They use a method to generate medical case-based questions by involving domain experts who identify possible diagnoses and related conditions. They build templates and generate questions per template, using information from experts to distinguish between diagnoses. For example, they identified six possible diagnoses for postoperative fever, set possible values, and assembled questions based on these conditions. Each template is specific to a sign or symptom, and questions generated from the same template can substitute for exams. Most work is done manually.	The AIG process generated 1248 multiple-choice items for diagnosing complications with postoperative fever. The 1248 items were produced in a total of 6 h across three stages: Stage 1 (3 h), Stage 2 (2 h), and Stage 3 (1 h)	Medical case-based questions.
Khodeir et al. (2014) [37]	Rule-based approaches: • They generate diagnostic questions using Bayesian network knowledge representation, but do not include standard patient demographics and histories. The most probable diagnosis is unclear, especially when two diseases are related to two symptoms.	There is a significant improvement in the approximation accuracy of the student model. The student model's ability to estimate or predict outcomes or behaviors is enhanced by 40%. Additionally, the paper mentions a 35% reduction in the number of assessing questions needed when adapted generated questions are utilized.	Medical case-based diagnostic questions.
Leo et al. (2019) [6]	Template-based approach: • Elsevier MCQ generator (EMCQG) is an MCQG system based on Elsevier Merged Medical Taxonomy-OWL (EMMeT-OWL) ontology, using built-in templates to generate unique, varying-difficulty questions based on EMMeT-OWL's classes, relations, and annotations.	The study generated 3,407,493 questions using an approach implemented by EMCQG.	Clinical dataset from EMMeT

The primary research gaps noted in Table 7 references are [6,26,34,36,37]: Automatically generated questions have a more simple composition as compared to manually created questions, typically consisting of only two stem entities and restricting the level of cognitive complexity. These questions are typically employed to assess the ability of learners to retrieve acquired knowledge, such as recalling definitions. The absence of advanced cognitive processes, such as the application of knowledge to novel scenarios, the analysis of knowledge, and the exercise of judgment, is a notable constraint in numerous educational programs. Although basic recall questions have their benefits, creating comprehensive tests requires a variety of questions that exhibit different patterns and cognitive complexity. It is crucial to have a wider variety of questions in the development of AI in education to accommodate various learning styles and cognitive abilities.

6.3. MCQs Generation Using Artificial Intelligence

Artificial intelligence (AI) is transforming the area of automated MCQ generation from text. With the help of AI tools, educational platforms can generate scenarios that mirror real-world medical situations, providing students and professionals with hands-on experiences and tests.

6.3.1. Machine Learning Approaches

Machine learning (ML) is a subset of artificial intelligence that enables machines to learn from data without being explicitly programmed [38]. It encompasses various approaches such as supervised, unsupervised, and reinforcement learning, allowing machines to improve their performance based on past experiences. Machine learning improves the process of generating MCQs through various technologies. These technologies surpass simple automation; they empower systems to acquire knowledge from current datasets, adjust, and gradually enhance the quality of generated questions.

Supervised Learning in MCQ Generation

Supervised learning (SL) has potential applications in the generation of case-based MCQs in medical education, although the direct application of SL in this domain is still emerging. SL algorithms, such as decision trees, random forests, and Naive Bayes, have been explored in the context of text classification and question generation in other domains, with promising implications for medical training. Several studies have successfully utilized supervised learning algorithms to recognize medical patterns, which could be applied to case-based MCQ generation [39,40]. For example, SL models have been shown to effectively classify complex medical data into hierarchical structures that simplify diagnosis [20]. Similarly, decision-tree SL models have demonstrated high accuracy in identifying key clinical features from patient data, improving diagnostic outcomes [41]. These approaches illustrate how SL can be leveraged to identify the critical elements necessary for generating case-based MCQs that assess clinical reasoning. The integration of patient cases, symptoms, diagnoses, and treatment data into structured formats has the potential to train SL models to produce accurate and clinically relevant questions, but further research and examples are needed to illustrate the direct use of SL for this purpose. Thus, while the theoretical foundation exists, future work must provide empirical evidence and detailed methodologies to show how SL can be used effectively in the medical domain for MCQ generation.

- **Recognizing Patterns and Learning:** The main goal of SL is to recognize patterns in medical cases and how questions are formulated based on those cases [42]. For example, an algorithm learns to generate questions from patient scenarios by recognizing specific clinical patterns, such as symptoms or test results that match known disease profiles [20]. Using supervised learning, the system is trained on labeled medical cases with the correct diagnoses, allowing it to learn patterns in the data and generate similar questions that assess a learner's ability to recognize symptoms and apply clinical knowledge. In this context, pattern recognition questions (PRQs) focus on identifying diseases from clinical patterns, requiring the examinee to match presented symptoms and test results with known disease patterns to make a diagnosis. This approach helps simplify complex decision-making by breaking down medical reasoning into smaller, manageable parts, thereby improving diagnosis or classification. In a study by Swe (2019), the decision tree supervised learning model is used to recognize medical patterns through a hierarchical structure [41]. The algorithm splits data into branches based on feature values (such as symptoms or test results), leading to a decision or classification at the tree's leaves. This structure simplifies complex decision-making by breaking it into smaller, more manageable parts, helping the model learn and identify key patterns in medical data for improved diagnosis or classification.

- **Extracting Features:** Feature extraction serves as the phase in learning algorithms, where crucial characteristics are derived from the training data [43]. In the context of MCQs, these key elements include scenarios, diagnostic criteria, treatment options and potential patient outcomes. This process plays a role in understanding the nuances of presenting cases and dealing with complexities when crafting questions. For example, decision tree and random forest algorithms are effective for feature extraction due to their inherent ability to perform feature selection during the model training process [41]. Random forest is an ensemble learning method primarily used for classification and regression tasks. It operates by constructing multiple decision trees during training and outputting the mode of the classes (for classification) or mean prediction (for regression) of the individual trees. These algorithms split data based on the most informative features, which helps in identifying and prioritizing features that contribute significantly to the decision-making process.
- **Training and Testing:** In the study by Yuan et al. (2017), a two-phase model is proposed to generate questions using supervised learning and reinforcement learning techniques [44]. Initially, the model is trained with teacher forcing to ensure it learns the correct sequence of outputs by maximizing the likelihood of ground-truth sequences. It is later fine-tuned with policy gradient reinforcement learning to allow improvements on sequences not encountered during training. The model's effectiveness is evaluated using the SQuAD dataset, and this approach can be adapted to generate medical MCQs by incorporating clinical datasets for relevance and accuracy in medical education.

Supervised learning has played a role in transforming how case-based MCQs are created [39]. By harnessing its potential, educators can develop high quality MCQs that serve as assessment tools while also enhancing deep learning and understanding in medical education. Continued progress and enhancements in this field will undoubtedly enhance the efficiency and scope of learning in generating content.

Unsupervised Learning in MCQ Generation

Unsupervised learning (UL), an approach of machine learning, holds the promise to transform the creation of multiple-choice questions based on cases. Unlike supervised learning, unsupervised learning (UL) operates without the need for labeled datasets. Instead, UL reveals underlying patterns and connections in data [45]. This characteristic is particularly beneficial for crafting innovative MCQs that can adapt to the changing landscape of medical knowledge. UL algorithms analyze literature, case studies and clinical information to generate MCQs by leveraging their ability to identify patterns and trends.

- **Extraction of Features from Unlabeled Data:** UL excels at extracting features from content. The system autonomously identifies concepts such as disease symptoms, diagnostic methods or treatment options and incorporates them into relevant inquiries [45]. This feature is particularly useful, for covering an array of topics and ensuring comprehensive educational resources. This process indeed exemplifies unsupervised learning as it involves discovering patterns and features in unlabeled data without predefined labels or categories.
- **Discovery of Patterns and Clustering:** The discovery of patterns and clustering, particularly through algorithms like hierarchical clustering, plays a crucial role in unsupervised learning (UL) for generating MCQs in medical education [46]. These algorithms group similar data, aiding in the identification of both common and rare data for question creation. Additionally, unsupervised information extraction (IE) techniques provide advantages by identifying important semantic relations between concepts from unannotated texts, without relying on pre-defined rules or patterns.

This approach enhances the flexibility and quality of MCQ generation, making it especially useful in contexts where manually annotated data are unavailable or costly.

- **Anomalies Detection:** One appealing aspect of unsupervised learning (UL) is its ability to identify anomalies [47]. In the context of MCQ generation, this entails pinpointing situations. These instances can serve as the foundation for creating demanding MCQs that assess a student's competence in managing scenarios.

A study outlines a framework that uses hierarchical clustering in unsupervised learning to generate medical question–answer pairs [48]. This approach integrates unsupervised key phrase detection and a multi-pass decoder, promoting diversity and validity in the generated pairs. The hierarchical Conditional Variational Autoencoder (CVAE) captures phrase-level relationships to maintain coherence. By incorporating both structured and unstructured knowledge, this method enhances the quality of medical questions, making hierarchical clustering a valuable tool for automatically structuring and diversifying medical MCQ generation.

Another study presents the Medical Topic discovery and Query (MedTQ) generation framework, which employs a hierarchical K-Means UL algorithm to uncover topics and generate queries from biomedical ontologies [49]. This technique uses a level-by-level optimization approach, ensuring strong associations within topics. It can be adapted for MCQ generation by identifying critical concepts and their relationships in medical datasets, thus facilitating the creation of meaningful and relevant multiple-choice questions.

Unsupervised learning provides a method for developing MCQs centered on scenarios. By utilizing its capability to detect patterns in data and adjust to information, UL can greatly improve the availability of current and challenging materials in the medical field. As it advances, its impact on the evolution of education is set to broaden and become increasingly significant.

6.3.2. Deep Learning Approaches

Deep learning, a subset of machine learning characterized by the use of neural networks with multiple layers, proves to be highly beneficial in the field of medicine due to its ability to process and analyze large volumes of complex medical data [50]. Approaches like natural language processing (NLP) and generative adversarial networks (GANs) have been utilized to create educational materials, such as MCQs. These techniques harness volumes of literature and case studies to teach models how to generate relevant and thought-provoking queries. MCQ generation using deep learning techniques has shown significant potential in the medical field, leveraging neural networks to create high-quality, relevant questions based on large datasets [51]. Text generation models based on the TensorFlow architecture, an open-source platform developed by Google for machine learning and deep learning, leverage different algorithms to generate high-quality MCQs. TensorFlow facilitates the building and training of deep learning models by providing a comprehensive ecosystem of tools and libraries that support the development of scalable and efficient machine learning applications [52]. These approaches enable the automatic generation of MCQs based on medical case reports, providing valuable educational resources for medical researchers and healthcare providers. By integrating deep learning methodologies into medical case-based MCQ generation, the system can offer enhanced learning opportunities and diagnostic support in the medical domain.

A deep learning technique known as the encoder–decoder (sequence-to-sequence) architecture, introduced by Google, supports end-to-end learning for tasks involving sequential input and output data [25]. This makes it particularly well-suited for text processing and generation tasks. The model comprises two components: an encoder, which processes the input sequence through multiple layers of a recurrent neural network, and a decoder, which uses similar layers to generate the output sequence. A recurrent neural network (RNN)-based model is a type of sequence-to-sequence deep learning algorithm designed to process sequential data by maintaining a "memory" of previous inputs through hidden states. This memory allows the model to capture dependencies between data points in a sequence, making it effective for tasks in text generation. This architecture is particularly effective for tasks involving sequences, such as the automatic generation of MCQs. In the study [53], question–answer pairs are generated using a knowledge graph combined with an RNN-based model. Keywords are extracted from the graph and then used in a sequence-to-sequence RNN model to create questions. The encoder–decoder architecture uses a bi-directional RNN with 1000 neurons in the hidden layer. Training on 1 million questions from Wiki Answers, the framework consists of two modules: one extracts entity knowledge, and the other generates questions. The RNN model achieved higher BLEU (Evaluating Matric) scores compared to phrase-based machine translation and template-based methods.

Another approach is generative adversarial networks (GANs) that are composed of two models: a generator and a discriminator [54]. The generator creates synthetic data, while the discriminator evaluates whether the data are real or generated. GANs have been successful in generating content such as images and text. In the medical field, GANs can be used for auto-generating MCQs by generating synthetic patient cases and questions, while the discriminator evaluates the quality. This process can help create diverse, realistic MCQs to simulate clinical scenarios, improving medical education and assessment outcomes.

6.3.3. Natural Language Processing (NLP)

Advances in natural language processing (NLP), a subset of deep learning, have enabled machines to better understand and generate human-like text [55]. NLP techniques are essential to generative models, such as Bidirectional Encoder Representations From Transformers (BERT) and Generative Pretrained Transformer (GPT), which are commonly applied in the automatic generation of medical MCQs [56]. These models excel at handling complex text generation tasks, including interpreting medical case studies and creating realistic questions. By leveraging transformer-based models, NLP has enhanced the development of MCQs. The following expanded explanation details how NLP streamlines each stage of generating MCQs in the field:

Medical Case Scenario Analysis and Extraction of Key Information

Medical case scenarios contain a wealth of data and complex medical terminology [6]. Utilizing NLP methods, these scenarios are broken down to reveal aspects like symptoms, patient history, diagnosis details and treatment results.

- **Entity Recognition:** Advanced NLP models perform Named Entity Recognition (NER) to identify and classify specific entities (e.g., diseases, drugs, symptoms) within the text [6]. This process helps in categorizing the information that can later be used to construct the stem of an MCQ.
- **Contextual Understanding:** NLP techniques used in generative models, like BERT, excel in interpreting the context of medical text [57]. For instance, BERT can differentiate between multiple meanings of the word "cold" (virus or temperature) based on the surrounding context. Its bidirectional architecture allows it to understand text from both left-to-right and right-to-left, which is essential for accurate medical text generation. By using these capabilities, BERT enhances the quality of automatically generated medical MCQs, ensuring contextually appropriate and accurate question formation.

Creation of Consistent, Authentic MCQ Prompts

After gathering all the details, it is crucial to convert them into MCQ prompts. This goes beyond making changes to the language; it involves incorporating knowledge into a framework that effectively evaluates understanding [6].

- **Scenario Simulation:** NLP can be used to simulate clinical scenarios that are realistic and relevant to the curriculum [58]. This involves creatively integrating different pieces of extracted information to form a scenario that mirrors real-life clinical situations.
- **Generation of Content:** Content generation involves the use of tools like GPT, which have been trained on extensive text data to produce coherent and contextually relevant

text [58]. Through natural language processing (NLP), these models generate content by predicting the next word in a sequence based on the preceding words. This capability allows for the construction of statements that adhere to typical medical evaluation standards while maintaining flexibility in expression. By leveraging the predictive power of NLP, these tools can generate questions and educational content that are both accurate and reflective of common medical scenarios, thereby enhancing the quality and relevance of the material.

Generation of Plausible Incorrect Options (Distractors)

Distractors are the incorrect options provided in an MCQ that are designed to be reasonable and challenging, without being misleading, in order to test the student's knowledge. Evaluating the learner's comprehension and capacity to apply knowledge critically is of utmost importance.

- **Similarity and Distinction of Semantic Content:** Techniques like semantic analysis enable NLP models to provide distractors that possess both similarity with and distinction from the correct response [59]. For instance, in a question regarding a treatment, the distractor options may consist of different pharmaceuticals that are suitable for related problems but are not suitable for the specific case specified in the question.
- **Analysis of Error Patterns:** NLP can evaluate patterns in students' misinterpretations of comparable instances, enabling the creation of distractors that mirror prevalent mistakes in the medical domain [60].

6.3.4. MCQ Generation Using AI Transformers

The use of AI transformers to generate medical case-based MCQs demonstrates a fusion of artificial intelligence and medical education [56]. Transformers, which belong to a category of deep learning models, have greatly revolutionized the field of NLP with their ability to effectively analyze data, understand context, and produce coherent text output. Their application in crafting MCQs is enhancing the way educators develop materials, resulting in more engaging, pertinent and challenging assessment tools.

AI transformers employ self-attention mechanisms, which allow the model to weigh the importance of different words in a sentence relative to each other [61]. This technique enables the processing and understanding of large amounts of data simultaneously, rather than sequentially, enhancing the model's ability to capture contextual relationships and dependencies within the text. This capability allows people to grasp the nuances of medical case stories and extract details that can be used to formulate multiple-choice questions. The various functions of AI transformers in generating MCQs are as follows:

- **Understanding Context**: Transformers can understand the shades of meaning in terms and concepts, setting them apart from other models in a case study setting [62]. This deep understanding allows for the creation of MCQs that are clinically precise and also intricately linked to context assessing students' capacity to apply medical knowledge in challenging real-world situations.
- **Data Scalability:** Experts can significantly enhance the capabilities of transformers by training them using a diverse mix of textbooks, journals, and case studies. This method allows transformers to capture a broader and more comprehensive range of information compared to traditional approaches like rule-based systems and long short-term memory (LSTM) networks, which often struggle with complex and varied medical data [63]. By leveraging this training, transformers can generate MCQs across different medical fields and specialties, offering flexible and scalable solutions to meet the growing demand for high-quality educational materials.
- **Innovative Question Formulation:** Transformers have sophisticated NLP capabilities that enable them to generate innovative and creative questions and answers [64]. Automated systems could generate subtle distractors (incorrect answers) that closely resemble typical misunderstandings or mistakes in clinical reasoning. This is achieved by training the models on large datasets that include examples of common errors and

misconceptions in medical practice. By analyzing these patterns, the models can produce distractors that are both plausible and challenging. This approach improves the instructional significance of the MCQs by encouraging students to engage in critical and discriminative thinking. One common method to evaluate the effectiveness of distractors is through item analysis, which involves statistical techniques to examine how test-takers respond to each option [65]. This can help identify which distractors are working well and which are not. In addition, distractors are reviewed by subject matter experts [66]. These experts can assess whether the distractors are plausible and relevant to the question being asked, ensuring they align with the guidelines of standardized exams like the United States Medical Licensing Examination (USMLE) and Comprehensive Osteopathic Medical Licensing Examination (COMLEX) of the United States. Moreover, analyzing how students interact with the distractors can provide valuable insights. If a distractor is never chosen, it may be too obviously incorrect. Conversely, if a distractor is chosen too frequently, it might be misleading or too similar to the correct answer. Tracking the number of times each distractor is selected can help identify which ones are effective and which need revision. Also, observing when and how often students engage with the practice quizzes can also indicate distractor effectiveness. For instance, if students frequently access quizzes and attempt questions multiple times, it suggests that the distractors are challenging enough to encourage repeated practice, which is a sign of their effectiveness. Another way to assess these distractors is collecting feedback from students about the distractors, which can provide direct insights into their effectiveness. Students can report if they found certain distractors confusing or too easy to eliminate, which can help in refining the questions to better assess their knowledge. Also, using statistical methods to analyze the performance of distractors can provide objective data. For example, item analysis techniques like the discrimination index can measure how well a distractor differentiates between high-performing and low-performing students. A good distractor should be more likely chosen by students who do not know the correct answer.

Although the utilization of transformers in the generation of MCQs shows potential, there are still obstacles to overcome, such as guaranteeing the clinical relevance and accuracy of the generated questions and avoiding the introduction of cognitive biases, such as confirmation bias (favoring information that confirms existing beliefs), availability bias (relying on easily accessible information), and anchoring bias (overemphasizing initial information) [67]. These biases in training data can lead to biased outputs in AI-generated content, which could mislead students or misrepresent medical knowledge. Therefore, careful evaluation and validation processes are essential to identify and mitigate these biases, ensuring that the generated questions are fair, unbiased, and educationally valuable. Ongoing collaboration among AI developers, medical educators, and clinicians is crucial to tackle these difficulties and enhance the technology [68]. The future of utilizing AI transformers in medical education is focused on advancing integration by combining multimodal data, such as clinical imaging, into the development of MCQs. Additionally, there is a growing interest in researching interactive, AI-driven simulations to enhance learning experiences with greater dynamism.

AI transformers are significantly enhancing medical education by utilizing their advanced capabilities to generate case-based MCQs. These transformers play a crucial role in helping medical students develop critical thinking and clinical reasoning skills by creating questions that are rich in context, clinically relevant, and educationally valuable [69]. This approach ensures that the MCQs are aligned with real-world medical scenarios, providing students with an effective and comprehensive learning experience. As this technology continues to advance, it holds great promise for further improving and customizing medical education. The future of AI-powered tools in training the next generation of healthcare practitioners looks bright, with these innovations poised to play a vital role in their education and preparation.

6.3.5. Case-Based MCQs Based on AI Applications

Various studies have investigated the utilization of artificial intelligence (AI) and machine learning to generate MCQs in the medical field. A study conducted a comparison between MCQs generated by the language model ChatGPT and those created by human examiners for medical graduate examinations [70]. In the training process, ChatGPT was trained using a vast amount of text data from the internet, which included books, articles, and websites, to understand and generate human-like text based on the input it received. This training helps the AI learn language patterns, grammar, and context. ChatGPT was specifically tasked with generating 50 MCQs using two standard undergraduate medical textbooks, Harrison's and Bailey & Love's, as references mentioned in the study. The AI was given prompts related to medical topics, and it created questions based on the information from these textbooks. To have a direct comparison between the questions generated by the AI and those created by experienced human educators, two university professors also created 50 MCQs using the same medical textbooks. All 100 MCQs generated by both ChatGPT and humans were then randomized and sent to five independent international assessors. These assessors evaluated the questions based on five criteria: appropriateness, clarity and specificity, relevance, discriminative power of alternatives, and suitability for medical graduate examinations. The assessors used a standardized assessment score to rate each question. The study found that ChatGPT-generated questions were comparable in quality to those created by humans, except in the relevance domain where AI was slightly inferior. The AI-generated questions also showed a wider range of scores, indicating variability in quality. One significant finding was that ChatGPT took only 20 min and 25 s to generate 50 questions, whereas the human examiners took a total of 211 min and 33 s, highlighting the efficiency of AI in generating educational content.

Another study was conducted in the generation of MCQs using ChatGPT based on specific scenarios [58]. ChatGPT was trained using a large dataset of text from the internet, which included books, articles, and websites, to understand and generate human-like text based on the input it received. The researchers created a detailed prompt for ChatGPT, which included specific instructions and context to generate case-based MCQs efficiently, reducing the need for extensive input from subject matter experts. The prompt incorporated cognitive and item models, which are frameworks that define the structure and content of the questions, ensuring that the generated MCQs were relevant and of high quality. The prompt included detailed medical scenarios and contexts, which guided ChatGPT to create realistic and contextually appropriate questions that assess higher-order thinking skills in medical students. By using a well-structured template, the prompt ensured that ChatGPT could generate a wide variety of questions, covering different topics and difficulty levels, which is essential for comprehensive medical education assessments. These prompts included specific instructions and examples to help the model produce high-quality questions that assess higher-order skills of medical students. After training, the model's output was tested and validated by subject matter experts to ensure the generated MCQs were accurate, relevant, and challenging. This step is crucial to ensure the quality and reliability of the questions before they are used in educational settings. The results highlight the potential of ChatGPT in medical education and encourage further research to explore and optimize its use, marking the beginning of the artificial intelligence era in this field.

Another investigation evaluated the suitability of AI models such as ChatGPT, Google Bard, and Microsoft Bing in producing MCQs that require reasoning for undergraduate medical students [71]. These models were trained using large datasets of text from the internet, which helped them understand and generate human-like text based on the input they received. These models use advanced algorithms to predict the next word in a sentence, allowing them to create coherent and contextually relevant responses. In this study, two physiologists selected specific competencies from the National Medical Commission of India's physiology curriculum, and then, a third physiologist prompted the AI models to generate five MCQs for each competency. This process was repeated for each of the

11 modules in the curriculum. ChatGPT generated the most valid MCQs, with a median validity score of 3 (on a scale of 0–3), while Bard and Bing had slightly lower validity scores, indicating that ChatGPT's questions were more accurate and relevant. However, ChatGPT's questions were also the least difficult, with a median difficulty score of 1. The reasoning ability required to answer the MCQs was rated similarly across all three AI models, with no significant difference, indicating that none of the models could generate questions that required a high level of subject understanding. This suggests that while AI can create basic questions, it struggles with more complex, reasoning-based queries. The study used statistical methods like the Kruskal–Wallis test to compare the distribution of scores and Cohen's Kappa to assess the agreement between the two raters, ensuring the reliability of the results. This rigorous evaluation helps in understanding the current capabilities and limitations of AI in generating educational content.

Another research paper introduced a novel approach to generating medical text responses using a bidirectional long short-term memory (Bi-LSTM) and neural network technology, focusing on the medical aesthetics domain, specifically double eyelid surgery questions and answers [72]. Bi-LSTMs are particularly useful for tasks that involve understanding context from both past and future data points in a sequence. The study utilized TensorFlow's Keras (high-level neural network API) to implement a sequential model, highlighting the embedding layer's role in processing textual data and the LSTM layer's contribution to maintaining contextual consistency in generated text. A unique dataset comprising queries and answers related to double eyelid surgery was used to train the model, demonstrating its capability to generate accurate medical text responses. The paper discusses the model's strengths, including its flexibility and ease of operation, while also acknowledging limitations such as its linear structure and inability to implement complex neural network topologies. Model evaluation metrics focused on accuracy, with the model showing promising results in generating coherent and contextually relevant text responses, suggesting potential for broader applications in medical intelligent question and answer systems.

7. Evaluation and Validation of Automatically Generated MCQs

Assessing and confirming the accuracy and quality of automatically generated medical MCQs is essential for ensuring their educational integrity and efficacy [17,52,53]. This evaluation process aims to ensure that the generated questions are accurate in terms of medical content and appropriately aligned with the intended difficulty level and instructional objectives. While the process helps enhance the quality and relevance of the questions, it does not provide a guarantee of accuracy without further validation and refinement. A comprehensive review and verification process ensures that MCQs align with goals accurately assessing student understanding and remain impartial. Additionally, these steps play a role in improving AI algorithms by ensuring the generated questions are relevant within their context and meet standards. In the realm of medical education, where the quality of learning directly influences future clinical performance, meticulous validation of MCQs is critical to preparing healthcare professionals with the skills necessary for effective patient care.

In evaluating the accuracy of created MCQs, various methods are typically utilized:

- **Precision:**

Precision is essential for assessing the quality of medical case-based MCQs. It ensures that the questions are accurate, relevant, and aligned with current medical knowledge, including correct terminology, diagnoses, and treatments [73]. Precision is critical in preventing errors that could mislead learners and propagate incorrect practices.

Precision, recall, and F1 score are crucial metrics for evaluating the accuracy of generated MCQs [74]. Precision measures the proportion of correctly generated questions among those identified as correct, ensuring that false positives are minimized. Recall evaluates the system's ability to identify all relevant questions, thus reducing false negatives by capturing all potential correct questions. The F1 score, as the harmonic mean of precision and recall, provides a balanced view, especially useful when there is an uneven class distribution or when one metric is prioritized over the other. Together, these metrics ensure that the generated MCQs are not only accurate but also comprehensive, covering all necessary aspects of the topic, which is essential for creating effective and reliable assessments.

- **Difficulty Score:**

Measuring the level of challenge when crafting MCQs is important to ensure they effectively assess learners. Difficulty is commonly assessed using models like item response theory (IRT) [75,76]. The Rasch model, for example, assesses question difficulty by analyzing responses from a large group of test takers and calculates the probability of a correct answer $(P(X=1))$ using the formula $(X=1)=\frac{e^{\theta-b}}{1+e^{\theta-b}}$, where θ is the examinee's ability and b is the item's difficulty. This model assumes that as a test taker's ability exceeds the question's difficulty, and their likelihood of answering correctly increases, thereby helping educators tailor questions to meet learning objectives [73,77]. Additionally, the 3-parameter logistic model in IRT considers discrimination (a), difficulty (b), and guessing (c) to refine the difficulty evaluation. The formula $P(\theta)=c+\frac{1-c}{1+e^{a(\theta-b)}}$ incorporates these parameters, where a indicates how well a question differentiates between skill levels, b represents the skill level for a 50% chance of correctness, and c accounts for guessing.

Using models like item response theory (IRT), questions can be specifically designed to target certain difficulty levels and discrimination parameters, which ensures a balanced assessment that can effectively differentiate between students' abilities [75,76]. IRT models also provide a detailed analysis of student responses, helping to identify which questions are too easy, too difficult, or ineffective at distinguishing between varying skill levels. This analysis guides educators in adjusting their question sets to better align with desired learning outcomes. As a result, these models ensure that the questions are both challenging and fair, offering a reliable measure of student learning and helping educators refine assessments to more accurately reflect students' progress.

In a study [78], AI-generated medical MCQs were evaluated for difficulty using expert judgment and psychometric analysis. Psychometric analysis is a statistical method used to evaluate the quality and effectiveness of test items by measuring factors like difficulty, discrimination (explained in next section), and reliability, ensuring the accuracy and fairness of assessments. Experts rated 80% of the items as easy and 20% as moderately difficult, but psychometric analysis showed that 90% of the items were moderately difficult based on student performance. A correlation between difficulty and discrimination indices was found, indicating that item difficulty relates to how well the questions differentiate between high and low performers. This highlights the effectiveness of using both expert and statistical evaluations in determining question difficulty. Another study [79] evaluated the difficulty score of generated MCQs using the difficulty index, which ranged from 0.3 to 0.9 for most items, indicating moderate difficulty. This range indicates that the questions were of moderate difficulty, meaning they were neither too easy nor too hard for the students.

- **Discrimination Index:**

The discrimination index in MCQs assesses how well a question can differentiate between students who excel and those who struggle, relative to a given population of students [80]. A high discrimination index indicates that a question effectively distinguishes individuals with a strong grasp of the subject from those who do not. To calculate this index, test takers are typically divided into two groups: achievers and low achievers, usually comprising the top 27% and bottom 27% of the population, respectively. The formula employed is as follows: Discrimination Index = (Number of accurate responses in the topmost group − Number of accurate responses in the lowermost group)/(Number of individuals in one group).

The range for this index varies from −1 to +1, with values suggesting differentiation. A positive index shows that a higher percentage of high-performing students answer the question correctly, whereas a negative index indicates potential issues with the question. This index is an empirical measure and requires conducting actual MCQ field tests with students to derive accurate values. In the context of generated MCQs, the discrimination index is crucial for validating question effectiveness and ensuring that the assessments can accurately differentiate between varying levels of student performance.

A study mentioned in the previous section [78] showed that the discrimination index of the generated medical MCQs ranged from 0.77 to 0.15. Of the items, 50% had excellent discrimination, 30% had good discrimination, 10% had poor discrimination, and 10% were non-discriminating.

- **Significance (Relevance):**

Relevance is an aspect in generating multiple-choice questions within medical education [6]. It pertains to how the questions correspond to the intended learning goals and the current standards of practice. It is essential for MCQs not just to cover content but to accurately reflect the intricate nature and depth of subjects in a way that suits the learner's educational level. One method to measure relevance is to engage a group of specialists in evaluating the questions. This assessment would determine how well each question fits with the curriculum or specific educational goals. A practical way to execute this idea would be to create a relevance scale that spans from 'not relevant' to 'very relevant'. Then, have experts evaluate each question accordingly. The relevance score for every multiple-choice question could be calculated by averaging all the scores, resulting in a relevance rating. Additionally, examining student performance on these questions using analysis could offer insights into their effectiveness, enhancing our understanding of their significance. For example, if students consistently excel on questions that experts consider relevant, it could confirm the efficacy of aligning the content.

The process of evaluating auto-generated MCQs typically involves experts in the field conducting assessments to verify their accuracy and relevance [81]. These experts utilize established guidelines and their knowledge to carefully examine the questions. This evaluation process often includes cross-referencing the MCQs with published literature, academic papers, or medical guidelines to ensure they align with current standards and knowledge. Additionally, statistical techniques, such as item analysis, can be used to assess the efficiency and consistency of these questions by studying student responses. This approach guarantees that the created MCQs accurately reflect the intended content and are suitable for their educational objectives. For example, studies have shown that auto-generated questions, when evaluated using these rigorous methods, can achieve a level of relevance and effectiveness comparable to human-composed questions.

- **Guessing Factor:**

The guessing factor in MCQs pertains to the likelihood of selecting an answer through guessing [76,82,83]. This factor is particularly significant when there are multiple answer options, as it increases the chances of making an educated guess. In a typical MCQ with four options, the probability of guessing the correct answer is 0.25, representing a 25% chance. However, more sophisticated approaches, such as item response theory (IRT) models, adjust this probability by analyzing how often students who perform poorly manage to answer correctly, potentially through guessing. This adjustment provides a more accurate indication of a question's ability to assess genuine comprehension rather than random luck.

When generating MCQs through automatic question generation (AQG) systems, it is important to consider the impact of guessing on student performance [35]. While AQG systems do not inherently incorporate item response theory (IRT) models, information from IRT-based analysis can be used to refine questions by identifying those that may be prone to guessing. During the validation phase, item analysis helps in adjusting or discarding questions that do not adequately differentiate between students who understand the material and those who might guess correctly. This process ensures that MCQs accurately assess learning outcomes while minimizing the effect of random guessing.

- **Feedback analysis:**

Feedback analysis is a critical method for evaluating the quality and efficacy of MCQs, particularly in the field of medical education [84]. This process involves collecting and analyzing input from both students and instructors regarding the clarity, relevance, and complexity of the MCQs. Quantitative methods, such as surveys or questionnaires, allow participants to rate various features of the MCQs on a numerical scale. Qualitative methods, including open-ended responses, provide deeper insights into specific issues. The gathered information can then be analyzed statistically to identify trends and significant issues, such as questions that are frequently misunderstood or perceived as too easy or too difficult.

For AQG systems, feedback analysis is essential to ensure the generated questions meet educational standards and learning objectives. Specific implications for AQG include the need for continuous refinement of algorithms based on feedback to improve question quality. Studies have shown that incorporating feedback analysis into the AQG process can significantly enhance the relevance and effectiveness of generated questions. For instance, Kurdi et al. (2020) demonstrated that using student feedback to iteratively improve question generation algorithms leads to better alignment with curriculum goals and improved student performance [2]. Also, they highlighted the importance of feedback loops in adaptive learning environments, where automatically generated questions are continuously refined based on learner responses. By consistently seeking input and implementing enhancements based on feedback, AQG systems can maintain high-quality question banks that effectively assess and enhance student learning [85]. This iterative process ensures that the questions not only align with the curriculum but also address the real-time needs and challenges faced by students.

- **ROUGE metric**

The Recall-Oriented Understudy for Gisting Evaluation (ROUGE) metric is utilized in natural language processing (NLP) to measure the quality of generated text by comparing it to a reference text in tasks such as summarization and machine translation [12,86]. ROUGE measures how similar the generated text is to the reference text, focusing on recall. This measure calculates the overlap of n-grams (sequences of words) between the reference text and the generated text. The basic calculation for ROUGE N is as follows: ROUGE N = (Number of shared n-grams in the reference and generated text)/(Total number of n-grams in the reference text).

For evaluating the generation of case-based MCQs, ROUGE can be used to assess the similarity of the generated questions to a set of preexisting human-created "gold-standard" MCQs [62,81]. However, this approach requires a substantial corpus of preexisting MCQs covering the same content to provide meaningful comparisons. While ROUGE can be useful during the system development phase to refine and validate the question generation algorithms, it cannot be used for evaluating new questions over entirely new materials.

- **BLEU metric**

The Bilingual Evaluation Understudy (BLEU) metric is commonly used in natural language processing to assess the quality of text in machine translation by comparing the generated text to reference translations [6,62,63]. BLEU measures the accuracy of the text based on how many n-grams from the generated text match those in the reference text [87]. The BLEU score is calculated by dividing the number of correct n-grams in the generated text by the number of n-grams present, with adjustments for brevity.

For MCQs, BLEU can be applied similarly to evaluate the quality of generated questions by comparing them to a set of standard or reference questions. This method provides information on the quality and relevance of the generated MCQs when compared to high-quality, human-created questions. However, like ROUGE, BLEU requires a significant number of preexisting MCQs on the same content to be effective. This limits its applicability for evaluating new questions over new materials but makes it valuable during the development and refinement of AQG systems.

For automatic question generation (AQG), using ROUGE and BLEU metrics involves comparing generated MCQs to a preexisting corpus of high-quality MCQs to ensure similarity in structure and content [4,81]. This process helps in refining AQG algorithms and ensuring that the generated questions meet the desired standards of quality and relevance. However, it is crucial to note that these metrics are best suited for system development rather than for the ongoing evaluation of new questions covering novel material. Therefore, additional evaluation methods, such as expert review and student feedback, are necessary to ensure the efficacy and appropriateness of newly generated MCQs.

- **Unweighted Kappa metric**

The unweighted Kappa, also known as Cohen's Kappa, is a statistical measure used to evaluate the level of agreement between two raters, accounting for agreement occurring by chance [88].In the context of evaluating generated MCQs, unweighted Kappa assesses the consistency between two experts' ratings on the quality, relevance, or appropriateness of the questions.

This involves creating a contingency table summarizing the ratings, where the observed agreement Po is calculated by dividing the number of questions both experts rated the same by the total number of questions, and the expected agreement Po is calculated based on the probability of chance agreement. The Kappa value is then derived using the formula: $= (Po - Pe)/(1 - Pe)$, where values range from 1 (perfect agreement) to 0 (agreement by chance) to negative values (less agreement than by chance). For instance, if two experts evaluate 100 MCQs and both rate 40 questions as acceptable and 40 as unacceptable, while disagreeing on the remaining 20 questions, the Kappa value would indicate good agreement (κ = 0.6). This measure provides a nuanced assessment of inter-rater reliability, essential for ensuring the quality and consistency of MCQs in educational settings.

- **Kruskal–Wallis Test**

The Kruskal–Wallis test is a non-parametric method used to evaluate the quality of generated MCQs by comparing ratings across three or more independent groups [69,89]. This test is particularly useful in educational research, as it does not assume a normal distribution of data, making it suitable for the ordinal data typically gathered from rating scales. For instance, in evaluating MCQs, researchers might gather ratings on aspects such as clarity, relevance, and difficulty from different groups of medical experts or students. The Kruskal–Wallis test ranks these ratings and calculates the H statistic to determine if there are significant differences in the medians among the groups. This helps identify whether certain groups consistently rate the questions higher or lower than others, providing valuable insights into the consistency and acceptance of the generated questions. Using the Kruskal–Wallis test ensures a robust statistical framework for assessing the effectiveness and fairness of automated MCQ generation systems, thereby enhancing the development of educational assessment tools.

8. Comparative Analysis and Evaluation of Automatic Case-Based MCQ Generation Applications

To provide a comprehensive overview of the progress and effectiveness of various automatic medical case-based MCQ generation methods, Table 8 summarizes recent applications, some of them discussed in the previous sections. This table includes key details such as the techniques used, datasets, performance metrics, evaluation metrics used and key findings. By offering a clear comparison, Table 8 serves as a valuable guide for future research and practical applications in the field.

Table 8. Summary analysis of recent studies on automatic case-based MCQ generation.

Study	Technique Used	Dataset	Performance	Evaluation Metrics Used	Key Findings
Leo et al. (2019) [6]	ontology-based approach	Clinical dataset from EMMeT	Application generated over 3 million questions across four physician specialties and conducted a user study involving 15 medical experts to evaluate the approach. The evaluation revealed that 129 questions (30%) were deemed appropriate for exam use by both experts, while an additional 216 questions (50%) were considered suitable by at least one expert.	Unweighted Kappa, Feedback analysis, Difficulty Score	The key findings of the study show that the ontology-based approach is effective in generating high-quality, complex MCQs that are suitable for medical education and assessment.
Huang et al. (2022) [72]	Bi-LSTM (Bidirectional Long Short-Term Memory), neural network technology	The dataset comprises queries and answers related to double eyelid surgery	This overview indicates that the application's performance is assessed through its optimization strategy, accuracy, loss rate, and user interaction capabilities, aiming to provide effective and reliable medical information.	Accuracy, Loss rate	The study does not explicitly detail the key findings or the exact figures for model accuracy and loss rate, but it emphasizes the importance of these metrics in evaluating the model's performance. The use of a specific medical dataset suggests a focused approach to improving AI-driven medical consultations.
Cheung et al. (2023) [70]	LLM (ChatGPT)	Two standard undergraduate medical textbooks (Harrison's, and Bailey & Love's)	ChatGPT was able to produce 50 questions in 20 min and 25 s, significantly faster than the 211 min and 33 s required by human examiners for the same number of questions. However, in the relevance domain, AI-generated questions scored slightly lower than human-generated ones.	Appropriateness, clarity and specificity, relevance	The study found no significant difference in the overall quality of questions between those generated by AI and humans, except in the relevance domain where AI was slightly inferior. AI-generated questions showed a wider range of scores, indicating variability in quality compared to the more consistent human-generated questions.
Y. S. Kıyak (2023) [58]	LLM (ChatGPT)	Not Specified	The ChatGPT prompt introduced in the paper can generate a large number of high-quality case-based multiple-choice questions (MCQs) quickly, which significantly reduces the effort required by subject matter experts in medical education.	Relevance, Feedback analysis	The paper finds that the ChatGPT prompt can generate a large number of high-quality case-based MCQs efficiently, significantly reducing the effort required by subject matter experts.
Agarwal et al. (2023) [71]	LLM (ChatGPT, Bard, and Bing)	Large datasets of text from the internet.	ChatGPT generated 110 MCQs, Bard generated 110 MCQs, and Bing generated 100 MCQs, as it failed to generate questions for two competencies.	Kruskal–Wallis Test, Unweighted Kappa	ChatGPT produced the most valid MCQs with a median validity score of 3, while Bard and Bing had slightly lower validity scores, indicating that ChatGPT's questions were more accurate and relevant. ChatGPT's questions were the least difficult, with a median difficulty score of 1, whereas Bard and Bing had slightly higher difficulty scores. The reasoning ability required to answer the MCQs was rated similarly across all three AI models, with no significant difference, indicating that none of the models could generate questions that required a high level of subject understanding.
Kiyak et al. (2024) [90]	LLM (ChatGPT)	Prompts published in medical education literature. Source not specified.	The performance of the case-based MCQ generator is enhanced by its ability to produce contextually relevant and high-quality MCQs efficiently, surpassing the capabilities of the standard ChatGPT	Relevance, Efficiency	The custom ChatGPT significantly streamlines the MCQ creation process by eliminating the need for manual prompt input, making it easier for medical educators to generate questions.

Table 8 provides a comparative analysis of different studies and techniques used for generating MCQs in the medical field. It covers a range of methodologies, from ontology-based approaches to advanced language models like ChatGPT, and examines their effectiveness in generating high-quality MCQs suitable for medical education. Key aspects evaluated include the datasets used, performance metrics such as accuracy, relevance, and feedback analysis, and the overall effectiveness of the approaches. For instance, Leo et al. (2019) demonstrated the effectiveness of an ontology-based approach in generating complex MCQs, while Huang et al. (2022) highlighted the importance of accuracy and loss rate in neural network models [6,72]. Similarly, studies by Cheung et al. (2023) and Kiyak (2023) emphasized the rapid generation of MCQs by ChatGPT, though with some variability in quality compared to human-generated questions [58,70].

The findings suggest that AI-driven approaches, particularly those using large language models like ChatGPT, can significantly reduce the time and effort required to generate MCQs while maintaining quality comparable to that of human-generated content. However, certain challenges, such as variability in question relevance and the need for precise evaluation metrics, were noted. ChatGPT, for example, was found to be slightly inferior in the relevance domain but excelled in efficiency, making it a valuable tool for medical educators. The table highlights the ongoing advancements in AI for medical education and the potential for these technologies to streamline the MCQ creation process.

9. Practical Implementation and Educational Impact

The auto-generation of MCQs in medical education has been explored in various educational settings, demonstrating significant impacts on student learning and assessment. One real-world example is the implementation of AI-driven systems to generate MCQs from medical textbooks and lecture notes, as discussed by Moore et al. [91]. This approach has been shown to enhance the efficiency of question creation, allowing educators to focus more on teaching and less on administrative tasks. In another study, Indran et al. explored the use of natural language processing (NLP) techniques to automatically generate MCQs from medical literature [92]. This method not only increased the volume of available assessment items but also ensured that questions were aligned with current medical knowledge, thereby improving the relevance and quality of assessments. The impact on student learning was notable, as students reported improved engagement and understanding of complex medical concepts due to the diverse range of questions generated. Berman et al. highlighted the use of auto-generated MCQs in formative assessments, which provided immediate feedback to students [66]. This feedback loop was crucial in helping students identify their strengths and weaknesses, leading to more targeted and effective study strategies. The study found that students who regularly engaged with these auto-generated MCQs performed better in summative assessments, indicating a positive impact on learning outcomes.

Rezigalla's research focused on the integration of auto-generated MCQs in online learning platforms, which facilitated continuous assessment and self-paced learning [78]. This approach was particularly beneficial in remote learning environments, where traditional assessment methods were challenging to implement. The flexibility and accessibility of these platforms contributed to improved student satisfaction and learning efficiency. Finally, Kıyak and Emekli examined the role of auto-generated MCQs in reducing cognitive load for educators, allowing them to allocate more resources to personalized student support and curriculum development [93]. This shift not only enhanced the educational experience for students but also improved the overall quality of medical education programs. In summary, the auto-generation of MCQs in medical education has proven to be a valuable tool in enhancing student learning and assessment. By leveraging AI and NLP technologies, educational institutions can provide more comprehensive and adaptive learning experiences, ultimately leading to better educational outcomes.

10. Research Gaps and Limitations

Automated MCQ generation in the medical field presents several challenges and gaps, as highlighted by recent research. One significant issue is the complexity of medical knowledge, which requires sophisticated algorithms to ensure the accuracy and relevance of generated questions. The intricacy of medical terminology and the need for context-specific understanding often lead to difficulties in generating questions that are both meaningful and educationally valuable [6]. Another gap is the lack of comprehensive datasets that can be used to train models for MCQ generation. The medical field is vast, and existing datasets may not cover all necessary topics or may lack the depth required for high-quality question generation. This limitation can result in questions that do not adequately test the breadth of knowledge expected from medical professionals [94]. Moreover, the evaluation of generated MCQs poses a challenge. There is a need for robust validation methods to ensure that the questions are not only syntactically correct but also pedagogically sound. Current systems often rely on human experts for validation, which can be resource-intensive and may not scale well with the increasing demand for automated educational tools [95]. Additionally, the integration of automated MCQ systems into existing educational frameworks is not seamless. There are concerns about the adaptability of these systems to different curricula and the potential for them to reinforce outdated or incorrect information if not regularly updated with the latest medical research [48]. Finally, there is a gap in the personalization of MCQs. Current systems often lack the ability to tailor questions to the individual learning needs and progress of students, which is crucial for effective learning in the medical field. Personalized learning paths could enhance the educational value of automated MCQs, but this requires further development in adaptive learning technologies [95]. In summary, while automated MCQ generation holds promise for medical education, significant gaps remain in terms of complexity handling, dataset comprehensiveness, validation processes, integration with educational systems, and personalization capabilities. Addressing these gaps will be essential for the effective deployment of automated MCQ systems in medical education.

11. Potential Areas for Further Investigation (Future Research Directions)

The incorporation of AI and ML methods, in developing MCQs inspired by actual cases marks a notable advancement [72,85]. These tools allow for the creation of customized questions tailored to the user's knowledge level, learning preferences, and specific educational needs. For instance, customization can include adjusting the difficulty level of questions based on the learner's performance, targeting specific learning objectives or topics that the student needs to focus on, and providing varied question formats to enhance engagement.

Through the use of algorithms, these platforms can assess individual learning situations to ensure that the generated questions are both stimulating and relevant. For example, a student struggling with a particular concept can receive more questions on that topic, with varying levels of difficulty to aid comprehension. Similarly, advanced learners can be challenged with more complex questions that push their understanding further.

Leveraging AI offers an opportunity to streamline the question-making process, leading to increased productivity and better alignment with educational goals [96]. As a result, the development of MCQs has seen progress in enhancing both the quality and effectiveness of assessments. Future research could explore more sophisticated adaptive learning systems, the integration of real-time feedback mechanisms, and the development of domain-specific question generation models to further personalize learning experiences.

11.1. Development of Data Sources and Interoperability

Integrating diverse data sources, such as clinical cases, literature, and patient records, significantly enhances the realism of case-based questions by providing a more comprehensive view of medical scenarios, which is essential for helping students grasp the complexities of real-life medical practice [97]. Clinical cases enable students to see the practical application of theoretical knowledge, while literature ensures that the information is accurate, evidence-based, and up-to-date, thus making the learning experience more reliable. Additionally, patient records add a crucial personal and social dimension to these cases, helping students appreciate the human experience of medicine and understand the patient's perspective. By combining these data types, educators can create cases that more effectively simulate real-life medical practice, better preparing students for their future roles as healthcare providers. This approach also encourages critical thinking, improving students' problem-solving skills and clinical reasoning, ultimately leading to more holistic and empathetic healthcare professionals who are well-prepared for the complexities of modern medical practice.

11.2. AI–Human Content Collaboration

The integration of AI technology with domain-specific human expertise offers promising potential for developing case-based MCQs [56,69]. This approach leverages the efficiency and data-processing capabilities of AI to generate questions, which are then refined through the practical insights and expertise of professionals. While AI algorithms can rapidly analyze large datasets to formulate initial questions, the final content is subject to rigorous review and validation by human experts to ensure it meets high standards of accuracy, relevance, and contextual appropriateness. This collaborative method aims to enhance the quality of educational content by combining the strengths of AI—such as speed and data handling—with the nuanced understanding of human professionals. Although still in development, this approach is expected to address challenges related to the creation of high-quality educational materials, while also considering ethical standards and the specific needs of learners. The ongoing evolution of AI in this domain holds the potential to significantly improve the effectiveness of training programs in medical education.

11.3. Adaptive Learning Systems and Personalization

The incorporation of adaptable learning systems into the creation of case-based MCQs represents a significant advancement in the field of medical education technology [98]. These systems employ algorithms to evaluate learners' responses, identifying their strengths, weaknesses and learning patterns. This information enables the generation of tailored questions that match each student's skill level and rate of learning. Adaptive learning algorithms adjust the question's difficulty, providing a customized learning journey that enhances educational outcomes. This approach not only enhances knowledge retention and understanding, it also fosters a more engaging and effective learning environment. These technologies have the potential to be used for ongoing professional development of healthcare practitioners, which would help them stay up-to-date with the continuously changing medical domain [73]. Potential advancements in this field may involve a more extensive incorporation of educational frameworks and a wider implementation across diverse medical fields, hence enhancing the educational environment for healthcare practitioners.

11.4. Research and Collaborative Development

Collaboration in developing automated methods for generating case-based MCQs is essential to advancing educational tools in medical training [4,27]. This interdisciplinary effort brings together experts in artificial intelligence, medical education, and healthcare to create MCQs that are not only contextually relevant but that also accurately reflect clinical realities. The primary goal is to refine AI algorithms for question generation, ensuring they incorporate clinical expertise to enhance accuracy and relevance in medical contexts. This collaborative approach is crucial for adapting educational tools to meet the diverse and evolving needs of modern medical education, ensuring that MCQs maintain the highest standards of clinical and educational quality. As technology and medical knowledge advance, such collaboration is vital for driving innovation and improving the effectiveness of medical education.

11.5. NLP Developments

The advancements in natural language processing (NLP) are changing how case-based MCQs are created, leading to improvements in accuracy and relevance [32,33]. The use of NLP algorithms now enables an analysis of texts leading to a grasp of complex medical terms and contexts. This advancement allows for the generation of questions that closely resemble real-world scenarios. The future outlook for NLP in MCQ development seems positive, with the potential for improvements in understanding and language models that can replicate the comprehension of literature. These advancements could result in learning experiences where questions cater to individual learning needs and also capture the nuances and depth of medical knowledge. The ongoing advancements in NLP technology are expected to facilitate its integration into tools ultimately creating learning platforms. This integration holds the potential to enhance the journey, for both aspiring healthcare professionals and experienced practitioners.

12. Conclusions

Automated case-based MCQ generation advances medical education technologies. Using machine learning, NLP, and ontology frameworks could simplify and speed up the process of MCQ creation. It guarantees that generated questions are standardized and simplify medical ideas. These questions improve learning outcomes depending on accuracy, relevance, difficulty, discrimination index, guess-ability factor, and feedback evaluation.

Future advances in machine learning, NLP, and ontological research, along with improved collaboration among educators, technologists, and healthcare experts, could improve education. We can create personalized learning experiences by adopting these technologies while maintaining goals, privacy, and ethics. These innovations may include tailored learning paths, interactive information delivery, and virtual and augmented reality simulations of medical settings for hands-on practice. In our comparative analysis of automatic case-based MCQ generation applications, it became evident that these technologies offer distinct advantages over commonly used methods, particularly in their ability to efficiently produce large volumes of high-quality questions. The analysis highlighted that while commonly used methods rely heavily on human effort and are time-consuming, automated systems can scale rapidly, maintaining consistency in question quality across different medical domains. However, the evaluation also pointed out challenges, such as the need for ongoing refinement of algorithms to better capture the nuances of medical knowledge and ensure the contextual relevance of the generated content.

Automated scenario-based MCQ generation changes medical knowledge teaching, assessment, and refinement, leading to an evolution in educational technology. This seamless combination of research and improvements foretells a future where technology-driven education generates trained and knowledgeable healthcare workers. In reaction to healthcare changes, education will become more efficient, accessible, and adaptable.

Author Contributions: Writing—original draft preparation, S.A.S.; review and editing, A.A.A., K.M., A.A. and H.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kumar, A.P.; Nayak, A.; Chaitanya, M.S.K.; Ghosh, K. A Novel Framework for the Generation of Multiple Choice Question Stems Using Semantic and Machine-Learning Techniques. *Int. J. Artif. Intell. Educ.* **2023**, *34*, 332–375. [CrossRef]
2. Kurdi, G.; Leo, J.; Parsia, B.; Sattler, U.; Al-Emari, S. A Systematic Review of Automatic Question Generation for Educational Purposes. *Int. J. Artif. Intell. Educ.* **2020**, *30*, 121–204. [CrossRef]

3. Falcão, F.; Costa, P.; Pêgo, J.M. Feasibility assurance: A review of automatic item generation in medical assessment. *Adv. Health Sci. Educ.* **2022**, *27*, 405–425. [CrossRef]
4. Al Shuriaqi, S.; Aal Abdulsalam, A.; Masters, K. Generation of Medical Case-Based Multiple-Choice Questions. *Int. Med. Educ.* **2023**, *3*, 12–22. [CrossRef]
5. Cohen Aubart, F.; Lhote, R.; Hertig, A.; Noel, N.; Costedoat-Chalumeau, N.; Cariou, A.; Meyer, G.; Cymbalista, F.; De Prost, N.; Pottier, P.; et al. Progressive clinical case-based multiple-choice questions: An innovative way to evaluate and rank undergraduate medical students. *Rev. Méd. Interne* **2021**, *42*, 302–309. [CrossRef]
6. Leo, J.; Kurdi, G.; Matentzoglu, N.; Parsia, B.; Sattler, U.; Forge, S.; Donato, G.; Dowling, W. Ontology-Based Generation of Medical, Multi-term MCQs. *Int. J. Artif. Intell. Educ.* **2019**, *29*, 145–188. [CrossRef]
7. Bansal, A.; Dubey, A.; Singh, V.K.; Goswami, B.; Kaushik, S. Comparison of traditional essay questions versus case based modified essay questions in biochemistry. *Biochem. Mol. Biol. Educ.* **2023**, *51*, 494–498. [CrossRef]
8. Gartmeier, M.; Pfurtscheller, T.; Hapfelmeier, A.; Grünewald, M.; Häusler, J.; Seidel, T.; Berberat, P.O. Teacher questions and student responses in case-based learning: Outcomes of a video study in medical education. *BMC Med. Educ.* **2019**, *19*, 455. [CrossRef]
9. Basuki, S.; Rizky, A.; Wicaksono, G.W. Case Based Reasioning (CBR) for Medical Question Answering System. *Kinet. Game Technol. Inf. Syst. Comput. Netw. Comput. Electron. Control* **2018**, *3*, 113–118. [CrossRef]
10. Majumder, M.; Saha, S.K. A System for Generating Multiple Choice Questions: With a Novel Approach for Sentence Selection. In Proceedings of the 2nd Workshop on Natural Language Processing Techniques for Educational Applications, Beijing, China, 31 July 2015; Association for Computational Linguistics: Stroudsburg, PA, USA, 2015; pp. 64–72. [CrossRef]
11. Madri, V.R.; Meruva, S. A comprehensive review on MCQ generation from text. *Multimed. Tools Appl.* **2023**, *82*, 39415–39434. [CrossRef]
12. Moon, H.; Yang, Y.; Shin, J.; Yu, H.; Lee, S.; Jeong, M.; Park, J.; Kim, M.; Choi, S. Evaluating the Knowledge Dependency of Questions. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, United Arab Emirates, 7–11 December 2022; pp. 10512–10526. [CrossRef]
13. Moore, S.; Costello, E.; Nguyen, H.A.; Stamper, J. An Automatic Question Usability Evaluation Toolkit. In *Artificial Intelligence in Education*; Olney, A.M., Chounta, I.-A., Liu, Z., Santos, O.C., Bittencourt, I.I., Eds.; Lecture Notes in Computer Science; Springer Nature: Cham, Switzerland, 2024; Volume 14830, pp. 31–46. [CrossRef]
14. Manoj, D.; Maria John, P. Natural language processing based question and answer generator. *Int. Adv. Res. J. Sci. Eng. Technol.* **2024**, *11*, 135–141. [CrossRef]
15. Dhanya, N.M.; Balaji, R.K.; Akash, S. AiXAM—AI assisted Online MCQ Generation Platform using Google T5 and Sense2Vec. In Proceedings of the 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS), Coimbatore, India, 23–25 February 2022; pp. 38–44. [CrossRef]
16. Maheen, F.; Asif, M.; Ahmad, H.; Ahmad, S.; Alturise, F.; Asiry, O.; Ghadi, Y.Y. Automatic computer science domain multiple-choice questions generation based on informative sentences. *PeerJ Comput. Sci.* **2022**, *8*, e1010. [CrossRef]
17. Paul, R.J.; Jamal, S.; Bejoy, S.; Daniel, R.J.; Aju, N. QGen: Automated Question Paper Generator. In Proceedings of the 2024 5th International Conference on Innovative Trends in Information Technology (ICITIIT), Kottayam, India, 15–16 March 2024; pp. 1–4. [CrossRef]
18. Ten Cate, O.; Custers, E.J.F.M.; Durning, S.J. (Eds.) *Principles and Practice of Case-Based Clinical Reasoning Education*; Innovation and Change in Professional Education; Springer International Publishing: Cham, Switzerland, 2018; Volume 15. [CrossRef]
19. Al-Rukban, M. Guidelines for the construction of multiple choice questions tests. *J. Fam. Community Med.* **2006**, *13*, 125. [CrossRef]
20. Freiwald, T.; Salimi, M.; Khaljani, E.; Harendza, S. Pattern recognition as a concept for multiple-choice questions in a national licensing exam. *BMC Med. Educ.* **2014**, *14*, 232. [CrossRef]
21. Family Medicine Modular Subject Exam—Content Outline. Available online: https://www.nbme.org/sites/default/files/2022-01/Family_Medicine_Sample_Items.pdf (accessed on 16 January 2024).
22. Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; et al. Gene Ontology: Tool for the unification of biology. *Nat. Genet.* **2000**, *25*, 25–29. [CrossRef]
23. El-Sappagh, S.; Franda, F.; Ali, F.; Kwak, K.-S. SNOMED CT standard ontology based on the ontology for general medical science. *BMC Med. Inform. Decis. Mak.* **2018**, *18*, 76. [CrossRef]
24. Bernasconi, A.; Masseroli, M. Biological and Medical Ontologies: Human Phenotype Ontology (HPO). In *Encyclopedia of Bioinformatics and Computational Biology*; Ranganathan, S., Gribskov, M., Nakai, K., Schönbach, C., Eds.; Academic Press: Oxford, UK, 2019; pp. 848–857. [CrossRef]
25. Mulla, N.; Gharpure, P. Automatic question generation: A review of methodologies, datasets, evaluation metrics, and applications. *Prog. Artif. Intell.* **2023**, *12*, 1–32. [CrossRef]
26. Wang, W.; Hao, T.; Liu, W. Automatic Question Generation for Learning Evaluation in Medicine. In *Advances in Web Based Learning—ICWL 2007*; Leung, H., Li, F., Lau, R., Li, Q., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2008; Volume 4823, pp. 242–251. [CrossRef]
27. Ladas, N.; Borchert, F.; Franz, S.; Rehberg, A.; Strauch, N.; Sommer, K.K.; Marschollek, M.; Gietzelt, M. Programming techniques for improving rule readability for rule-based information extraction natural language processing pipelines of unstructured and semi-structured medical texts. *Health Inform. J.* **2023**, *29*, 146045822311646. [CrossRef]

28. Xue, X.; Wu, Q.; Ye, M.; Lv, J. Efficient Ontology Meta-Matching Based on Interpolation Model Assisted Evolutionary Algorithm. *Mathematics* **2022**, *10*, 3212. [CrossRef]
29. Das, R.; Ray, A.; Mondal, S.; Das, D. A rule based question generation framework to deal with simple and complex sentences. In Proceedings of the 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Jaipur, India, 21–24 September 2016; pp. 542–548. [CrossRef]
30. Rao, P.R.; Jhawar, T.N.; Kachave, Y.A.; Hirlekar, V. Generating QA from Rule-based Algorithms. In Proceedings of the 2022 International Conference on Electronics and Renewable Systems (ICEARS), Tuticorin, India, 16–18 March 2022; pp. 1697–1703. [CrossRef]
31. Zhang, R.; Guo, J.; Chen, L.; Fan, Y.; Cheng, X. A Review on Question Generation from Natural Language Text. *ACM Trans. Inf. Syst.* **2022**, *40*, 1–43. [CrossRef]
32. Patil, P.M.; Bhavsar, R.P.; Pawar, B.V. A Review on Natural Language Processing based Automatic Question Generation. In Proceedings of the 2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS), Trichy, India, 24–26 November 2022; pp. 1–6. [CrossRef]
33. Mehta, P.K.; Jain, P.; Makwana, C.; Raut, D.C.M. Automated MCQ Generator using Natural Language Processing. *Int. Res. J. Eng. Technol.* **2021**, *8*, 2705–2710.
34. Karamanis, N.; Ha, L.A.; Mitkov, R. Generating Multiple-Choice Test Items from Medical Text: A Pilot Study. In Proceedings of the Fourth International Natural Language Generation Conference, Sydney, Australia, 15–16 July 2006; Association for Computational Linguistics: Stroudsburg, PA, USA, 2006; pp. 111–113.
35. Mitkov, R.; An Ha, L.; Karamanis, N. A computer-aided environment for generating multiple-choice test items. *Nat. Lang. Eng.* **2006**, *12*, 177–194. [CrossRef]
36. Gierl, M.J.; Lai, H.; Turner, S.R. Using automatic item generation to create multiple-choice test items. *Med. Educ.* **2012**, *46*, 757–765. [CrossRef]
37. Khodeir, N.; Wanas, N.; Darwish, N.; Hegazy, N. Bayesian based adaptive question generation technique. *J. Electr. Syst. Inf. Technol.* **2014**, *1*, 10–16. [CrossRef]
38. Mendonça, M.O.K.; Netto, S.L.; Diniz, P.S.R.; Theodoridis, S. Chapter 13—Machine learning: Review and trends. In *Signal Processing and Machine Learning Theory*; Diniz, P.S.R., Ed.; Academic Press: Cambridge, MA, USA, 2024; pp. 869–959. [CrossRef]
39. Ono, S.; Goto, T. Introduction to supervised machine learning in clinical epidemiology. *Ann. Clin. Epidemiol.* **2022**, *4*, 63–71. [CrossRef]
40. Uddin, S.; Khan, A.; Hossain, M.E.; Moni, M.A. Comparing different supervised machine learning algorithms for disease prediction. *BMC Med. Inform. Decis. Mak.* **2019**, *19*, 281. [CrossRef]
41. Swe, T.T. Analysis of Tree Based Supervised Learning Algorithms on Medical Data. *Int. J. Sci. Res. Publ.* **2019**, *9*, p8817. [CrossRef]
42. Mondal, N.; Lohia, M. Supervised Text Classification using Text Search. *arXiv* **2020**, arXiv:2011.13832.
43. Ahmadi, S.A.; Mehrshad, N.; Razavi, S.M. Supervised feature extraction method based on low-rank representation with preserving local pairwise constraints for hyperspectral images. *Signal Image Video Process.* **2019**, *13*, 583–590. [CrossRef]
44. Yuan, X.; Wang, T.; Gulcehre, C.; Sordoni, A.; Bachman, P.; Zhang, S.; Subramanian, S.; Trischler, A. Machine Comprehension by Text-to-Text Neural Question Generation. In Proceedings of the 2nd Workshop on Representation Learning for NLP, Vancouver, BC, Canada, 3 August 2017; Association for Computational Linguistics: Stroudsburg, PA, USA, 2017; pp. 15–25. [CrossRef]
45. Talukdar, J.; Singh, T.P.; Barman, B. Unsupervised Learning. In *Artificial Intelligence in Healthcare Industry*; Talukdar, J., Singh, T.P., Barman, B., Eds.; Springer Nature: Singapore, 2023; pp. 87–107. [CrossRef]
46. Afzal, N.; Mitkov, R. Automatic generation of multiple choice questions using dependency-based semantic relations. *Soft Comput.* **2014**, *18*, 1269–1281. [CrossRef]
47. Yousefpour, A.; Shishehbor, M.; Foumani, Z.Z.; Bostanabad, R. Unsupervised Anomaly Detection via Nonlinear Manifold Learning. *arXiv* **2023**, arXiv:2306.09441. [CrossRef]
48. Shen, S.; Li, Y.; Du, N.; Wu, X.; Xie, Y.; Ge, S.; Yang, T.; Wang, K.; Liang, X.; Fan, W. On the Generation of Medical Question-Answer Pairs. *arXiv* **2019**, arXiv:1811.00681. [CrossRef]
49. Shen, F.; Lee, Y. MedTQ: Dynamic Topic Discovery and Query Generation for Medical Ontologies. *arXiv* **2018**, arXiv:1802.03855.
50. Bas, A.; Topal, M.O.; Duman, C.; Van Heerden, I. A Brief History of Deep Learning-Based Text Generation. In Proceedings of the 2022 International Conference on Computer and Applications (ICCA), Cairo, Egypt, 20–22 December 2022; pp. 1–4. [CrossRef]
51. Hu, Y.; Han, G.; Liu, X.; Li, H.; Xing, L.; Gu, Y.; Zhou, Z.; Li, H. Design and Implementation of a Medical Question and Answer System Based on Deep Learning. *Math. Probl. Eng.* **2022**, *2022*, 1–6. [CrossRef]
52. Zou, H. AIADA: Accuracy Impact Assessment of Deprecated Python API Usages on Deep Learning Models. *J. Softw.* **2022**, *17*, 269–281. [CrossRef]
53. Reddy, S.; Raghu, D.; Khapra, M.M.; Joshi, S. Generating Natural Language Question-Answer Pairs from a Knowledge Graph Using a RNN Based Question Generation Model. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. Valencia, Spain, 3–7 April 2017; Association for Computational Linguistics: Stroudsburg, PA, USA, 2017; pp. 376–385. [CrossRef]
54. Mitra, N.K.; Chitra, E. Glimpses of the Use of Generative AI and ChatGPT in Medical Education. *Educ. Med. J.* **2024**, *16*, 155–164. [CrossRef]

55. He, X.; Nassar, I.; Kiros, J.; Haffari, G.; Norouzi, M. Generate, Annotate, and Learn: NLP with Synthetic Text. *Trans. Assoc. Comput. Linguist.* **2022**, *10*, 826–842. [CrossRef]
56. Biswas, D.; Nadipalli, S.; Sneha, B.; Gupta, D.; Amudha, J. Natural Question Generation using Transformers and Reinforcement Learning. In Proceedings of the 2022 OITS International Conference on Information Technology (OCIT), Bhubaneswar, India, 14–16 December 2022; pp. 283–288. [CrossRef]
57. Ferrando, J.; Gállego, G.I.; Tsiamas, I.; Costa-jussà, M.R. Explaining How Transformers Use Context to Build Predictions. *arXiv* **2023**, arXiv:2305.12535.
58. Kıyak, Y.S. A ChatGPT Prompt for Writing Case-Based Multiple-Choice Questions. *Rev. Esp. Educ. Méd.* **2023**, *4*, 98–103. [CrossRef]
59. Nemani, P.; Vollala, S. A Cognitive Study on Semantic Similarity Analysis of Large Corpora: A Transformer-based Approach. In Proceedings of the 2022 IEEE 19th India Council International Conference (INDICON), Kochi, India, 24–26 November 2022; pp. 1–6. [CrossRef]
60. Yunjiu, L.; Wei, W.; Zheng, Y. Artificial Intelligence-Generated and Human Expert-Designed Vocabulary Tests: A Comparative Study. *SAGE Open* **2022**, *12*, 215824402210821. [CrossRef]
61. Tay, Y.; Bahri, D.; Metzler, D.; Juan, D.-C.; Zhao, Z.; Zheng, C. Synthesizer: Rethinking Self-Attention in Transformer Models. *arXiv* **2021**, arXiv:2005.00743.
62. Miller, K. Comprehension of Contextual Semantics Across Clinical Healthcare Domains. In Proceedings of the 2022 IEEE 10th International Conference on Healthcare Informatics (ICHI), Rochester, MN, USA, 11–14 June 2022; pp. 479–480. [CrossRef]
63. Chandraju, A.V.; Gnanasigamani, L.J. Transformer-Based Abstract Generation of Medical Case Reports. *Int. J. Eng. Adv. Technol.* **2022**, *12*, 110–113. [CrossRef]
64. Rodriguez-Torrealba, R.; Garcia-Lopez, E.; Garcia-Cabot, A. End-to-End generation of Multiple-Choice questions using Text-to-Text transfer Transformer models. *Expert Syst. Appl.* **2022**, *208*, 118258. [CrossRef]
65. Rao, M.C.; Sreedhar, P.; Bhanurangarao, M.; Sujatha, G. Automatic Multiple-Choice Question and Answer (MCQA) Generation Using Deep Learning Model. In Proceedings of the 2nd International Conference on Cognitive and Intelligent Computing, Hyderabad, India, 27–28 December 2022; Kumar, A., Ghinea, G., Merugu, S., Eds.; Springer Nature: Singapore, 2023; pp. 1–8.
66. Berman, J.; McCoy, L.; Camarata, T. LLM-Generated Multiple Choice Practice Quizzes for Pre-Clinical Medical Students; Use and Validity. *Physiology* **2024**, *39*, 376. [CrossRef]
67. Moradi, M.; Samwald, M. Improving the robustness and accuracy of biomedical language models through adversarial training. *J. Biomed. Inform.* **2022**, *132*, 104114. [CrossRef]
68. Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; Galstyan, A. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* **2022**, *54*, 1–35. [CrossRef]
69. Denecke, K.; May, R.; Rivera-Romero, O. Transformer Models in Healthcare: A Survey and Thematic Analysis of Potentials, Shortcomings and Risks. *J. Med. Syst.* **2024**, *48*, 23. [CrossRef]
70. Cheung, B.H.H.; Lau, G.K.K.; Wong, G.T.C.; Lee, E.Y.P.; Kulkarni, D.; Seow, C.S.; Wong, R.; Co, M.T.H. *ChatGPT Versus Human in Generating Medical Graduate Exam Questions—An International Prospective Study*; Medical Education: Tokyo, Japan, 2023. [CrossRef]
71. Agarwal, M.; Sharma, P.; Goswami, A. Analysing the Applicability of ChatGPT, Bard, and Bing to Generate Reasoning-Based Multiple-Choice Questions in Medical Physiology. *Cureus* **2023**, *15*, e40977. [CrossRef]
72. Huang, K.; Ji, F.; Lu, W.; Xiao, Y. Research on Text Generation of Medical Intelligent Question and Answer Based on Bi-LSTM and Neural Network Technology. In Proceedings of the 2022 IEEE/ACIS 22nd International Conference on Computer and Information Science (ICIS), Zhuhai, China, 26–28 June 2022; pp. 54–59. [CrossRef]
73. Sileo, D.; Uma, K.; Moens, M.-F. Generating Multiple-Choice Questions for Medical Question Answering with Distractors and Cue-Masking. *arXiv* **2023**, arXiv:2303.07069.
74. Sykes, B.; Simon, L.; Rabin, J. Unifying and Extending Precision Recall Metrics for Assessing Generative Models. *arXiv* **2024**, arXiv:2405.01611.
75. Embretson, S.E.; Reise, S.P. *Item Response Theory for Psychologists*; Lawrence Erlbaum Associates Publishers: Mahwah, NJ, USA, 2000; p. xi, 371. [CrossRef]
76. Isnawati, I.; Sriyati, S.; Agustin, R.R.; Supriyadi, S.; Kasi, Y.F.; Ismail, I. Analysis of Question Difficulty Levels Based on Science Process Skills Indicators Using the Rasch Model. *Tadris J. Kegur. Dan Ilmu Tarb.* **2024**, *9*, 31. [CrossRef]
77. Demaidi, M.N.; Gaber, M.M.; Filer, N. Evaluating the quality of the ontology-based auto-generated questions. *Smart Learn. Environ.* **2017**, *4*, 7. [CrossRef]
78. Rezigalla, A.A. AI in medical education: Uses of AI in construction type A MCQs. *BMC Med. Educ.* **2024**, *24*, 247. [CrossRef]
79. Alqahtani, S. Multiple choice questions as a tool for summative assessment in medical schools. *Bull. Egypt. Soc. Physiol. Sci.* **2024**, *44*, 29–38. [CrossRef]
80. Mahjabeen, W.; Alam, S.; Hassan, U.; Zafar, T.; Butt, R.; Konain, S.; Rizvi, M. Difficulty Index, Discrimination Index and Distractor Efficiency in Multiple Choice Questions. *Ann. PIMS.* **2017**, *13*, 310–315.
81. Kurdi, G.; Parsia, B.; Sattler, U. An Experimental Evaluation of Automatically Generated Multiple Choice Questions from Ontologies. In *OWL: Experiences and Directions—Reasoner Evaluation*; Dragoni, M., Poveda-Villalón, M., Jimenez-Ruiz, E., Eds.; Springer International Publishing: Cham, Switzerland, 2017; pp. 24–39.
82. Cooper, B.; Foy, J.M. Guessing in Multiple-choice Tests. *Med. Educ.* **1967**, *1*, 212–215. [CrossRef]

83. May, K. Book Review: Fundamentals of Item Response Theory Ronald K. Hambleton, H. Swaminathan, and H. Jane Rogers Newbury Park CA: Sage, 1991, 174 pp. *Appl. Psychol. Meas.* **1993**, *17*, 293–294. [CrossRef]
84. Rai, N.; Rai, N. Multiple choice questions: As formative assessment. *Int. J. Med. Biomed. Stud.* **2019**, *3*, 75–79. [CrossRef]
85. Das, B.; Majumder, M.; Phadikar, S.; Sekh, A.A. Automatic question generation and answer assessment: A survey. *Res. Pract. Technol. Enhanc. Learn.* **2021**, *16*, 5. [CrossRef]
86. Shaheer, S.; Hossain, I.; Sarna, S.N.; Kabir Mehedi, M.H.; Rasel, A.A. Evaluating Question generation models using QA systems and Semantic Textual Similarity. In Proceedings of the 2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 8–11 March 2023; pp. 431–435. [CrossRef]
87. Sellam, T.; Das, D.; Parikh, A.P. BLEURT: Learning Robust Metrics for Text Generation. *arXiv* **2020**, arXiv:2004.04696.
88. Mishra, S. Nitika Understanding the calculation of the kappa statistic: A measure of inter-observer reliability. *Int. J. Acad. Med.* **2016**, *2*, 217. [CrossRef]
89. Bobbitt, Z. Kruskal-Wallis Test: Definition, Formula, and Example. Available online: https://www.statology.org/kruskal-wallis-test/ (accessed on 16 January 2024).
90. Kıyak, Y.S.; Kononowicz, A.A. Case-based MCQ generator: A custom ChatGPT based on published prompts in the literature for automatic item generation. *Med. Teach.* **2024**, *46*, 1018–1020. [CrossRef]
91. Moore, S.; Schmucker, R.; Mitchell, T.; Stamper, J. Automated Generation and Tagging of Knowledge Components from Multiple-Choice Questions. In Proceedings of the Eleventh ACM Conference on Learning @ Scale, Atlanta, GA, USA, 18–20 July 2024; ACM: New York, NY, USA, 2024; pp. 122–133. [CrossRef]
92. Indran, I.R.; Paranthaman, P.; Gupta, N.; Mustafa, N. Twelve tips to leverage AI for efficient and effective medical question generation: A guide for educators using Chat GPT. *Med. Teach.* **2024**, *46*, 1021–1026. [CrossRef]
93. Kıyak, Y.S.; Emekli, E. ChatGPT prompts for generating multiple-choice questions in medical education and evidence on their validity: A literature review. *Postgrad. Med. J.* **2024**, qgae065. [CrossRef]
94. Murphy Lonergan, R.; Curry, J.; Dhas, K.; Simmons, B.I. Stratified Evaluation of GPT's Question Answering in Surgery Reveals Artificial Intelligence (AI) Knowledge Gaps. *Cureus* **2023**, *15*, e48788. [CrossRef]
95. Abdallah, A.; Kasem, M.; Hamada, M.A.; Sdeek, S. Automated Question-Answer Medical Model based on Deep Learning Technology. In Proceedings of the 6th International Conference on Engineering & MIS 2020, Almaty, Kazakhstan, 14–16 September 2020; ACM: New York, NY, USA, 2020; pp. 1–8. [CrossRef]
96. Ahamed, S.H.; Reddy, K.R.K.; Shoba, L.K. Enhancing Education with NLP-through AI-Enhanced Q&A Evaluation and Testing using Leveraging algorithms. In Proceedings of the 2024 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), Chennai, India, 9–10 May 2024; pp. 1–7. [CrossRef]
97. MacLeod, A.; Luong, V.; Cameron, P.; Burm, S.; Field, S.; Kits, O.; Miller, S.; Stewart, W.A. Case-Informed Learning in Medical Education: A Call for Ontological Fidelity. *Perspect. Med. Educ.* **2023**, *2*, 120–128. [CrossRef]
98. Pugh, D.; De Champlain, A.; Gierl, M.; Lai, H.; Touchie, C. Can automated item generation be used to develop high quality MCQs that assess application of knowledge? *Res. Pract. Technol. Enhanc. Learn.* **2020**, *15*, 12. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.