

A review of automatic item generation techniques leveraging large language models

Bin Tan^{1*}, Nour Armoush¹, Elisabetta Mazzullo¹, Okan Bulut², Mark J. Gierl²

¹University of Alberta, Faculty of Education, Measurement, Evaluation, and Data Science, Edmonton, Canada

²University of Alberta, Faculty of Education, Centre for Research in Applied Measurement and Evaluation, Edmonton, Canada

ARTICLE HISTORY

Received: Dec. 16, 2024

Accepted: Apr. 28, 2025

Keywords:

Automatic item generation,
Automatic question generation,
Large language models,
Generative artificial
intelligence,
Educational technology.

Abstract: This study reviews existing research on the use of large language models (LLMs) for automatic item generation (AIG). We performed a comprehensive literature search across seven research databases, selected studies based on predefined criteria, and summarized 60 relevant studies that employed LLMs in the AIG process. We identified the most commonly used LLMs in current AIG literature, their specific applications in the AIG process, and the characteristics of the generated items. We found that LLMs are flexible and effective in generating various types of items across different languages and subject domains. However, many studies have overlooked the quality of the generated items, indicating a lack of a solid educational foundation. Therefore, we share two suggestions to enhance the educational foundation for leveraging LLMs in AIG, advocating for interdisciplinary collaborations to exploit the utility and potential of LLMs.

1. INTRODUCTION

As a foundational component of the modern education system, assessments serve multiple purposes such as facilitating student learning, evaluating teaching outcomes, and informing educational policy (Black, 1998; Lee *et al.*, 2020). Advancements in educational technology, such as e-learning platforms (Granić *et al.*, 2022), adaptive testing (Yen *et al.*, 2012), and formative assessments (Dalby & Swan, 2019; Spector *et al.*, 2016), have increased the demand for high-quality assessment items. However, traditional item development methods are struggling to keep pace with this demand. For example, the cost of developing a single item for high-stakes assessments can reach up to \$2,000 (Rudner, 2010). Additionally, items for low-stakes assessments often suffer from low-quality content, thus failing to provide comprehensive or valuable feedback to students (Lim, 2019; Wylie & Lyon, 2015).

In response to the growing demand for high-quality items, educational measurement researchers first turned their attention toward a solution called automatic item generation (AIG). The goal of AIG is to generate large banks of high-quality items while reducing overall costs

*CONTACT: Bin TAN ✉ btan4@ualberta.ca 📍 Measurement, Evaluation, and Data Science, Faculty of Education, University of Alberta, 6-110 Education Centre North, 11210 87 Ave NW, Edmonton, AB T6G 2G5 CANADA

The copyright of the published article belongs to its author under CC BY 4.0 license. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>

(Gierl & Lai, 2012; Lai *et al.*, 2009). The template-based method proposed by Gierl *et al.* (2012) is the most widely used method for AIG, which involves a three-step process that requires the construction of cognitive models and item models (Gierl & Lai, 2012, 2016). This method has been successfully implemented in various educational practices (e.g., medical licensing examinations), demonstrating its viability and cost-effectiveness in creating high-quality items (Kosh *et al.*, 2018; Pugh *et al.*, 2020). Coincidentally, there is a related line of research in the domain of computer science, specifically focusing on automatic question generation (AQG) for educational purposes. Originally rooted in natural language processing (NLP) research for developing chatbots and conversational agents, AQG for educational purposes has evolved to share a similar goal with AIG: the automatic generation of a large number of assessment items using computer algorithms. Despite this shared goal, key terminological differences persist between the two fields. In educational measurement, the term “items” can be presented in diversified types to measure students' knowledge and skills, while “questions” typically have a much narrower format, usually referring to interrogative sentences ending with a question mark (Gierl *et al.*, 2021). Although an item may include a question, not all questions qualify as complete items. Despite this distinction, researchers of AQG for educational purposes often use the term “questions” broadly, even when referring to non-interrogative item formats.

The work conducted in AQG research for educational purposes has been summarized in two recent review studies (Ch & Saha, 2018; Kurdi *et al.*, 2020). Regrettably, in both reviews, there were no keywords of “automatic item generation” or even word variants of “items” in the search terms used. As a result, many significant publications in the AIG literature were not included in their review. Furthermore, educational measurement researchers have been less familiar with AQG due to differences in research interests, technical skill sets, and limited participation in NLP conferences. This gap is evident in the minimal overlap between the references and citations used in the two research domains. For example, nearly all references presented in the review paper of Kurdi *et al.* (2020) were published in computer science venues, while the references used in a review paper of AIG (Falcao *et al.*, 2022) were mostly in education, psychology, and testing journals. As a result, historically, researchers in both domains have had limited knowledge of each other's work or were not even aware of each other's existence. More importantly, the use of different terminology (e.g., “items” vs. “questions”) to describe and address similar problems has created communication barriers between the AIG and AQG research domains, hindering their advancement.

Marked as recent breakthroughs in generative artificial intelligence research, pre-trained language models (PLMs) and large language models (LLMs) have overcome the limitations of earlier autoregressive algorithms (e.g., rule-based systems) which relied on fixed architectures for rigid, sequential text prediction. Unlike traditional autoregressive algorithms, modern language models offer the distinct advantage of being retrainable and fine-tunable, enabling them to develop more sophisticated language comprehension and produce more contextually appropriate assessment items (e.g., Vu & Van Nguyen, 2022). These language models focus specifically on text generation but have also demonstrated their flexibility and extraordinary performance in understanding the context of language and in various language processing tasks (Ackerman & Balyan, 2023). The versatility and potential application of the models has significantly sparked the interest of both educational measurement researchers and applied NLP researchers in employing them as an innovative methodology for automating the generation of a large number of items. Technically, LLMs are essentially scaled-up versions of PLMs, characterized by an increased number of parameters and larger training data sizes. However, there is no strict cutoff value that differentiates the number of parameters of PLMs from those of LLMs (Zhao *et al.*, 2023). Therefore, in the remaining sections of this review, the term “LLMs” is used to refer to both PLMs and LLMs, as it is a more commonly recognized term.

As AIG and AQG researchers are increasingly sharing the same goal of generating questions for educational purposes and are also adopting similar methodologies such as the use of LLMs,

there is substantial decrease in the distinctions between them. Therefore, it has become crucial to raise awareness of both AIG and AQG research and map the literature of both fields to gain a comprehensive understanding of the utility and potential of LLMs. The purpose of this review is not only to identify what has been achieved in the literature regarding the use of LLMs for AIG but also to explore the knowledge gaps between the two research domains. Additionally, we aim to provide suggestions for future research to advance automated item generation practices in both fields. To ensure consistent terminology, we will use “AIG” to refer to both AIG and AQG for educational purposes in this paper, as the term “item” is more inclusive than “question” (Gierl *et al.*, 2021). However, it is important to note that these terms have distinct historical origins and development trends, but they are used interchangeably here due to their similar research objectives and the focus on leveraging LLMs in this review. To map the current state of research on leveraging LLMs for AIG, our review is guided by the following research questions (RQs):

RQ1: What are the most used LLMs for AIG, and how do their underlying architectures and features differ?

RQ2: In what specific ways are LLMs most frequently used, and how do the LLMs’ features influence their use in the AIG process?

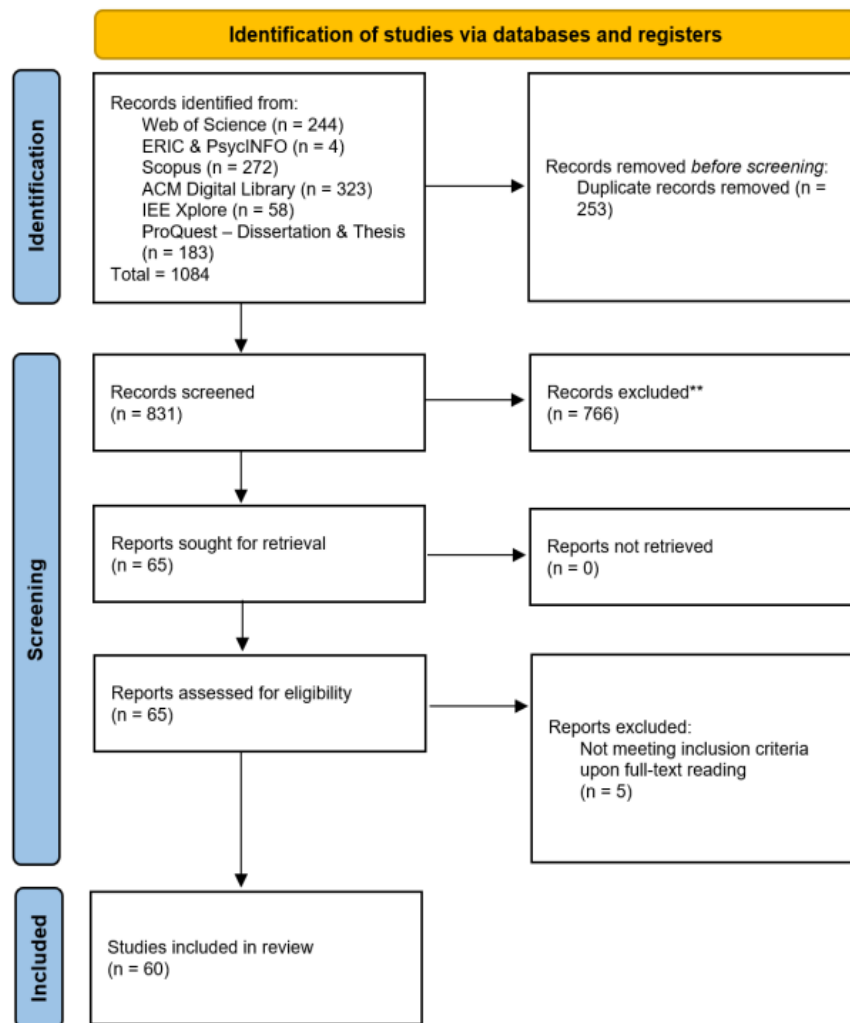
RQ3: What types of items were generated in the reviewed studies in terms of item type, subject domain, and language? Can LLMs generate valid, reliable, and high-quality items?

2. METHOD

2.1. Study Search and Selection

The study search and selection process are displayed in [Figure 1](#). Keyword searches were conducted in seven research databases, which were selected for their comprehensive coverage of literature on AIG (Alsubait, 2015) or relevance to LLMs. The searches used two keyword components: AIG and LLMs, connected by an "AND" operator, as detailed in Table 1. Our search targeted full texts, with no restrictions on publication types in the search strategy, and was limited to publications after 2018, marking the emergence of pre-trained language models (Zhao *et al.*, 2023). The last search was conducted on August 1st, 2023, which yielded 831 unique study records from these research databases.

After the study search, two reviewers independently screened the titles, abstracts, and keywords of the 831 identified studies, using predefined inclusion criteria for selection. We only retained studies with full texts in English and focused exclusively on empirical studies, excluding editorials, commentaries, and position papers. Given that question generation has applications in other domains like chatbots and customer service, we confined our selection to studies in the education field. This focus was achieved by identifying education-related words or phrases in the title, abstracts, or keywords. Additionally, we selected publications that focused on item generation, excluding other forms of text generation, such as the creation of reading passages without items generated. After the initial screening, we noted 47 disagreements in selection decisions, resulting in a high agreement rate of 93.44%. The disagreements primarily resulted from vague mentions of education-related phrases in abstracts from papers published in computer science venues. These disagreements were resolved through further discussion and consultation of the full texts of the studies, resulting in the selection of 65 studies for full-text review and information extraction. After a detailed review, 5 studies were excluded based on our criteria, leaving 60 studies in this review.

Figure 1. The flow diagram of search process.**Table 1.** Keywords used for study search.

Keywords related to AIG	Keywords related to LLMs
Item generation*, AIG, Question generation, Distractor generation, Test development, Item development, Generat* Item*, Generat* Question*	Large language model, Pre-trained language model, PLM, LLM, language model, BERT, GPT*, ChatGPT, T5, mT5, PanGu, T0, CodeGen, Tk-Instruct, UL2, OPT, NLLB, GLM, Flan-T5, BLOOM, mT0, Galactia, BLOOMZ, OPT- IML, LLaMA, CodeGeeX, Pythia, GShard, Codex, ERNIE, Jurassic-1, HyperCLOVA, FLAN, Yuan, Anthropic, WebGPT, Gopher, GlaM, LaMDA, MT-NLG, AlphaCode, InstructGPT, Chinchilla, PaLM, AlexaTM, Sparrow, WeLM, U-PaLM, Flan-U-PaLM, (Transformer AND Model)

Note: We focused our search on publications written in English and published in or after 2018. We conducted searches using titles, abstracts, and keywords. “*” represents variants of the word, such as plurals.

2.2. Information Extraction and Analysis

To address the proposed research questions, a coding framework was developed through an iterative process to extract information from the included studies. Initially, one reviewer proposed a preliminary coding framework and piloted it by extracting information from 20 of the included studies. Refinements to the coding framework, such as adding new coding

subsections, removing redundant or irrelevant ones, and defining coding categories within each subsection, were conducted. Corresponding to the RQs, the finalized coding framework comprised three key aspects: (1) the specific LLMs employed in each study, (2) their implementation methods categorized by role in the AIG process, and (3) the characteristics of generated items. Subsequently, with the coding framework, two reviewers independently coded all the studies after training. During this process, several ambiguities emerged, such as the distinctions between LLM applications (e.g., item filtering versus quality ranking for quality control purposes) and classification of LLM roles in multi-stage AIG processes (as LLMs might be employed at multiple stages). To resolve these ambiguities and any coding disagreements, the reviewers first discussed their interpretations, with a third reviewer consulted when consensus could not be reached. In the final analysis phase, one reviewer systematically quantified the studies according to each coding category and synthesized the findings. For instance, LLM applications were categorized into three distinct AIG process stages: pre-generation stage, item generation stage, and post-generation stage.

While conducting the information extraction, we realized some information was missing in the current literature for us to answer our research questions. Particularly, many studies did not report the measurement properties of the generated items. Therefore, we employed a simple search strategy to identify the severity of this issue – searching for keywords that are expected to appear in the reviewed studies. We developed a set of basic keywords that are related to the measurement properties of items as well as individuals who are often in charge of evaluating item quality such as content specialists and SMEs. Then, we built a simple information extractor using the Python programming language and used it to summarize the patterns of occurrences of keywords across the reviewed studies.

3. RESULTS

Our review indicated that there were two studies published in 2019, followed by an increase to five in 2020, 9 in 2021, a peak of 33 in 2022, and finally, eleven studies as of August 2023. The included studies can be found in [Appendix A](#). Over half of these studies ($n = 31$) were papers published at conferences, such as the *International Conference on Artificial Intelligence in Education* and the *European Conference on Technology Enhanced Learning*. Fourteen of the reviewed studies were journal articles, which span two broad research domains, with educational technology and assessment journals like *Education and Information Technologies* and *International Journal of Assessment Tools in Education*, and applied artificial intelligence journals such as *Frontiers in Artificial Intelligence* and *IEEE Access*. Moreover, there were non-refereed studies, including two master's theses, one doctorate dissertation, and 12 preprints. These avenues imply the multidisciplinary and developing nature of studies in this field.

3.1. RQ1: What Are the Most Used LLMs for AIG, and How Do Their Underlying Architectures and Features Differ?

We identified Google's Text-to-Text Transfer Transformer (T5; $n = 32$), Bidirectional Encoder Representations from Transformers (BERT; $n = 26$), and OpenAI's Generative Pre-Trained Transformer (GPT; $n = 19$) as three major base types of LLMs commonly used in AIG. In addition to these most common LLMs, we also found models that were only used once or a couple of times such as BART and PEGASUS. Because there were cases where more than one LLM was used in a study, we summarized the frequency of use by instance. That is, each study could be counted multiple times depending on how many LLMs they used.

Though all follow the transformer-learning paradigm, each type of LLM employs a distinct approach to language modeling, encompassing unique features and strengths for specific tasks. Due to the transfer learning paradigm, these LLMs often undergo further training with additional text data to enhance the base model to the desired task. This results in a variety of

variants and series. For a better understanding of their use in the AIG stages, the features and variants of each base type of LLM are outlined in the following subsections.

3.1.1. T5 variants

Introduced by Raffel *et al.* in 2020, the T5 model treats all language processing tasks as text-to-text conversions. That is, the model receives text input and directly generates text output. The dataset used to train this LLM was the C4 dataset which contains approximately 750GB of English texts sourced from the public Common Crawl web scrape (Raffel *et al.*, 2020). The T5 model is available in several sizes, such as T5-small, T5-base, and T5-large, each differing in the number of parameters. The different sizes of these models enable their application in varied contexts, accommodating diverse requirements in terms of computational resources and performance capabilities.

3.1.2. BERT variants

BERT (Devlin *et al.*, 2018) employs a bidirectional approach to understanding the context of words within a sentence. That is, the model analyzes texts from both left to right and right to left, which helps it gain a more comprehensive understanding of sentence contexts. BERT was originally trained using 3.3 billion words sourced from Wikipedia and BooksCorpus, based on two key training strategies: masked language model and next-sentence prediction. BERT can also be further trained with additional text data or training strategies, allowing it to exhibit differential performance in specific contexts. For example, multilingual BERT (Devlin *et al.*, 2019) are trained with additional text data from multiple languages to improve performance in multilingual understanding. The size of training datasets, training architecture, and the number of parameters can also lead to variants such as ALBERT (BERT with reduced parameters and sentence order prediction; Lan *et al.*, 2019), DistilBERT (BERT with distillation; Sanh *et al.*, 2019), RoBERTa (BERT without NSP but dynamic masking; Liu *et al.*, 2019), and XLNet (BERT with all tokens masked but in random order; Yang *et al.*, 2019).

3.1.3. GPT series

GPT (OpenAI, 2018) employs the next-word prediction strategy to generate words sequentially. That is, it predicts the next word in a sentence based on probabilities conditioned on both the original input it receives (i.e., the prompt) and the words it has previously generated. The series of GPT models has seen remarkable advancements over the last few years, with each iteration increasing in parameters and enhancing capabilities and complexity. In this series, GPT-1 contained 117 million parameters, GPT-2 expanded to 1.5 billion parameters, while GPT-3 further increased the scale to 175 billion parameters. At the time of this review, GPT-4 (Open AI, 2023) stood as the latest iteration, reflecting substantial improvements in language processing, reasoning, and multimodal functionality. Subsequent models have continued to advance the field, focusing not only on scale but also on real-world usability by optimizing for faster inference speeds, reduced computational costs, and greater affordability for end users. Furthermore, these models have become increasingly accessible, with intuitive platforms enabling seamless deployment for both developers and non-technical users.

Building upon the core models of such as GPT-4 and GPT-3, Open AI has developed specialized derivatives tailored for specific applications. For example, the well-known ChatGPT (Open AI, 2022) is specifically fine-tuned to engage in interactive conversational tasks. This fine-tuning involves encoding given human language inputs (prompts) into rich, contextual-embedded textual representations that the machine system can understand. Subsequently, it decodes these textual representations sequentially, conditioned on the original prompts and the previously generated words, to produce coherent and contextually appropriate responses in human languages. Similarly, Codex, built based on GPT-3, is trained to specialize in understanding and generating codes of programming languages (Open AI, 2021).

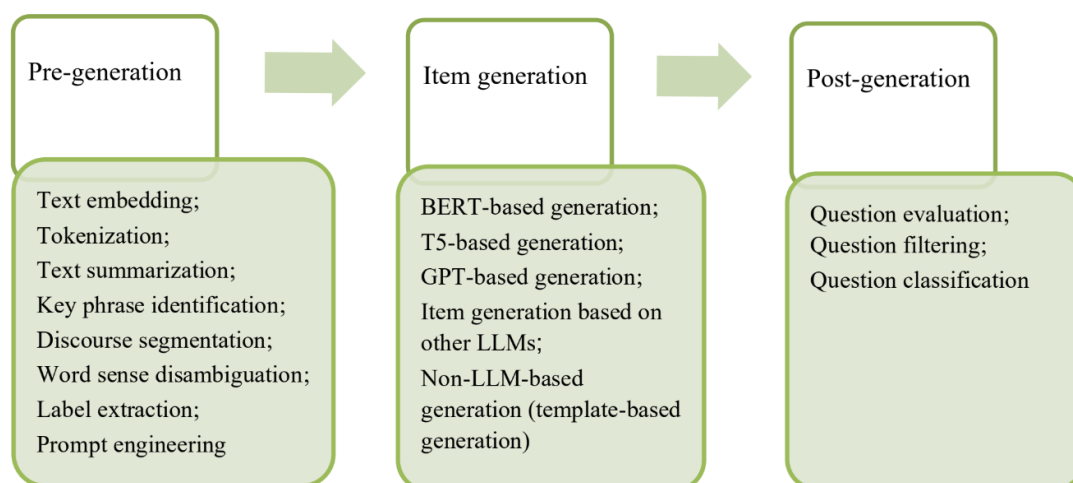
3.1.4. Summary of findings for RQ1

The most used base types of LLMs are T5, BERT, and GPT. These models exhibit differences in their training data sources, the size of their training sets, their training architectural frameworks, and their unique features.

3.2. RQ2: In What Specific Ways Are LLMs Most Frequently Used, and How Do the LLMs' Features Influence Their Use in the AIG Process?

As depicted in Figure 2, we summarized and synthesized the specific uses of LLMs into three stages: pre-generation, item generation, and post-generation. In the pre-generation stage, LLMs are used for preparation tasks before generating items such as cleaning, structuring, and understanding the original input text data which can be lengthy and covers various topics. This preparation is essential for the subsequent item generation stage to produce items that are contextually relevant, accurate, and coherent. In the item generation stage, items are created either directly by LLMs or through traditional methods like template-based approaches. Studies using template-based approaches were included because they employed LLMs in either the pre- or post-generation stages. After item generation is the post-generation stage, where items are evaluated and selected based on certain criteria such as quality and difficulty. With LLMs, this stage can be automated, resulting in filtered items as the final output of the AIG process. The majority of these studies ($n = 44$) employed LLMs in only one of the three AIG stages, 13 studies used LLMs in two stages, and 3 studies used LLMs across all three stages.

Figure 2. The specific use of LLMs in the AIG process.



3.2.1. Pre-generation stage

In the pre-generation stage, BERT was used 20 times, T5 was used 13 times, and GPT was used 9 times. We identified and categorized the following specific uses of LLMs: (1) text embedding ($n = 8$), (2) key phrase identification ($n = 8$), (3) text summarization ($n = 6$), (4) word tokenization ($n = 3$), (5) prompt engineering ($n = 2$), (6) word sense disambiguation ($n = 2$), (7) label extraction ($n = 1$), and (8) discourse segmentation ($n = 1$). These specific uses make up a toolbox for processing texts for subsequent tasks of generating desired items. The specific uses of LLMs were categorized into three themes, as described below.

The first theme is data transformation, which converts original texts into computationally tractable and manageable units. For example, word tokenization is the process of transforming texts into smaller, more manageable tokens. This task is often necessary before many subsequent preprocessing NLP tasks. LLM tokenizers have several advantages over traditional rule-based tokenizers, such as understanding context (Singh *et al.*, 2019). In addition, text embedding transforms text into numerical vectors, known as textual representations. These

numerical vectors can capture the semantic meaning of the original texts; texts with similar meanings are positioned closely in the embedding space. These numerical vectors are used and processed in later language processing tasks. For instance, they can be utilized for subsequent NLP tasks such as calculating the cosine similarity metric between items or distractors of multiple-choice questions to evaluate their semantic similarity (e.g., Min *et al.*, 2021).

The second theme is related to semantic processing, which involves understanding the context of the text data or specific sentences. For example, key phrase identification involves extracting important and relevant phrases from texts. This step can help to frame the focus of the item generation. For example, Tsai *et al.* (2021) employed BERT to extract keywords from an input textbook, which were then used to construct important complete sentences as preparation for the item generation stage. In addition, word sense disambiguation is the process of determining the actual meaning of a word with multiple meanings in a given sentence and context. This step is crucial for computers to interpret words correctly in context, ensuring that later generated items are contextually appropriate and unambiguous. Moreover, we found that discourse segmentation was performed in two studies. This task divides texts into coherent segments such as sentences, paragraphs, or topics, which helps to structure the text and generate items in a way that reflects the logical and semantic composition of the input texts. Furthermore, label extraction assigns labels to texts according to categories like reading difficulty, content domain, or question type. The extracted labels serve as control labels for creating targeted and relevant items. For example, Zhao *et al.* (2022) trained an LLM to extract one of seven question types from input texts, which was then used to generate questions matching the intended type. Lastly, text summarization aims to convey the main points of the original text while significantly reducing its length, thereby improving the efficiency of subsequent item-generation tasks (Malhar *et al.*, 2022).

The last theme is prompt engineering, which involves constructing commands or instructions that effectively communicate a task to LLMs. Some generative artificial intelligence such as GPT operate based on the prompt it receives. In AIG, the characteristics and quality of the prompt determine how LLMs generate items that meet specific assessment criteria, cater to various contexts, or assess across different domains. Therefore, researchers have used LLMs to generate effective prompts for subsequent tasks. For instance, in the study by Ghanem *et al.* (2022), T5 was trained on how to formulate questions and consider relevant aspects for subsequent item generation.

3.1.2. Item generation stage

We identified that 53 studies used LLMs during the item generation stage. Some studies used multiple LLMs, either to complement each other in different tasks or to compare and identify the best-performing models in the same task. Despite BERT and GPT-2 being introduced earlier than T5, the latter emerged as the most frequently employed LLM for item generation, being used 33 times. For example, Akyön *et al.* (2022) trained a variant of T5 (mT5) to first extract answers and then used these extracted answers as text input to generate questions as text output. GPT also comes as a popular option, having been used 15 times. GPT-variants predict the next word in a text sequence, enabling prompt-based generation. For instance, Wang *et al.* (2022) compared the performance of GPT-3 in item generation when using different prompts. BERT was used in 9 instances. Given BERT's characteristics, BERT variants were employed to predict masked tokens, creating specific item types such as fill-in-the-blank and cloze questions (e.g., Matsumori *et al.*, 2023). In addition, we identified instances where other LLMs, such as BART and PEGASUS, were used.

3.1.3. Post-generation stage

To ensure the quality and relevance of the generated items for a given educational context, LLMs can be used in the post-generation stage to evaluate, filter, or classify the previously generated items. Among them, evaluation was conducted by calculating specific criteria or

metrics. For instance, Jiao *et al.* (2023) utilized GPT-2 to calculate perplexity values, reflecting the fluency of the generated items. They also employed BERT-large for coherence evaluation and BertScore for assessing creativity by calculating semantic differences among items. Having the evaluation metrics calculated, filtering involves removing low-quality or irrelevant questions or distractors of multiple-choice questions that were generated by the LLM. For example, Offerijns *et al.* (2020) used GPT-2 to generate question-answer pairs and then used BERT to remove the questions that could not be answered or did not make sense. Lastly, classification refers to categorizing questions by type or topic. For example, Nguyen *et al.* (2022) first employed T5 to generate questions and then used GPT-3 to classify these questions based on their utility in learning specific topics.

3.1.4. Summary of findings for RQ2

We revealed distinct usage patterns of BERT, GPT, and T5 across the three AIG stages, reflecting the inherent features and strengths of these LLMs. As BERT excels at language understanding, it was predominantly utilized in the pre-generation stage. On the other hand, GPT was primarily used in the item-generation stage, highlighting its strength in text generation. T5 demonstrated its flexibility by being used comparably in all three AIG stages.

3.3. RQ3: What Types of Items Were Generated in the Reviewed Studies in Terms of Item Type, Subject Domain, and Language? Can LLMs Generate Valid, Reliable, and High-Quality Items?

3.3.1. Item type

We found 39 studies that generated constructed-response items; 29 studies generated selected-response items. Eight studies generated both types. The constructed-response items can be further categorized as Wh-questions, cloze questions, and Fill-in-the-blank questions, whereas selected-response items included True-False and multiple-choice questions, including their distractor generation.

3.3.2. Subject domain

Our review revealed that the majority of items generated primarily focus on two subject domains: language learning ($n = 24$) and general knowledge acquisition ($n = 17$). General knowledge items are often developed using general datasets, such as the Stanford Question Answering Dataset (SQuAD), which contains passages from Wikipedia and thus does not focus on a specific subject domain. Following closely behind are science-related disciplines ($n = 9$) such as agronomy, biology, chemistry, physics, and science history. In addition, computer science and mathematics education, had five and four studies on item generation, respectively. We also found studies addressing other subject domains, including social science, medicine, and even literary experiences such as fairy tales.

3.3.3. Language

Overall, we identified a total of 12 different languages in the items generated. This finding suggests the potential for generalizability and the wide linguistic context in which LLMs can be utilized for AIG. In most instances, items were generated in English ($n = 51$), but there were also instances of generation in other languages, including Arabic, Chinese, French, German, Hindi, Indonesian, Korean, Lao, Marathi, Spanish, Swedish, Turkish, and Vietnamese.

3.3.4. Data source

LLMs are typically developed for general language processing purposes and often require additional training to effectively perform specific tasks, such as generating context-relevant questions. Such additional training involves training LLMs on new datasets to adapt them to different tasks or content domains. In the reviewed studies, the most commonly utilized datasets for additional training have included SQuAD ($n = 24$), which encompasses a collection of passages with corresponding reading comprehension questions, and the ReAding

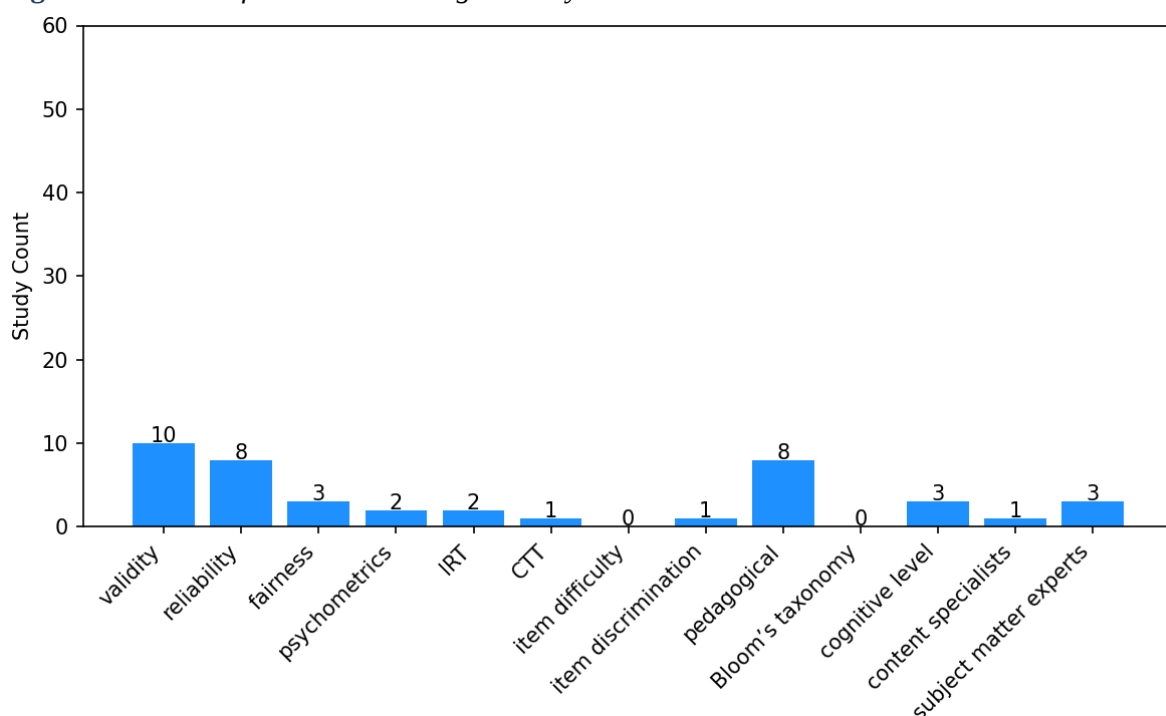
Comprehension dataset from Examinations (RACE) dataset ($n = 10$), which focuses on English language exams. Additionally, some researchers have crafted their self-collected datasets by aggregating educational materials. These materials consisted of open-access journal articles (e.g., von Davier 2019), questions extracted from online learning platforms or communities (e.g., Stack Overflow; Tsai *et al.*, 2021), LLM-generated texts (Bulut & Yildirim-Erbasli, 2022), teacher-created questions (Matsumori *et al.*, 2023), stories and fairy tales (e.g., Ghanem *et al.* 2022), knowledge maps (Aigo *et al.*, 2021), slides (Chughtai *et al.*, 2022), textbooks (e.g., Steuer *et al.* 2020), and other course materials (e.g., Gopal, 2022).

3.3.5. Item properties

As noted in the methodology section, we did not find many studies reporting the measurement properties of the generated items. Therefore, instead, we searched for keywords pertinent to measurement properties across the 60 reviewed studies. Figure 3 depicts the number of studies containing each keyword. Notably, only 10 out of the 60 studies mentioned “validity”. This was followed by “reliability” and “pedagogical”, which found their places in eight studies. Other keywords were used even less frequently in the reviewed studies. The infrequent occurrences of these keywords signal a concerning issue: the majority of the reviewed studies seem to neglect measurement properties of items when generating items for educational purposes, which potentially impacts the validity and reliability of the assessment results. Moreover, this could result in generating questions that are too simple and do not require higher cognitive thinking to answer, failing to meet the measurement or pedagogical purposes (e.g., delivering feedback or evaluating achievement).

We further extracted sentences containing the keywords. However, we found that many occurrences of these keywords were not related to the measurement properties of the generated items. For instance, most descriptions of “reliability” were in contexts other than the reliability of items or assessment results. They commonly referred to terms like “inter-annotator reliability” and “the reliability of the data collection”. Similarly, the instances of “fairness” were exclusively related to the fairness of experiments comparing model performance, rather than to the fairness of the assessment items themselves. Therefore, the issue of lacking sufficient consideration for the measurement properties of items may be even more severe than it appears in Figure 3.

Figure 3. Number of studies containing each keyword.



3.3.6. Summary of findings for RQ3

We found that LLMs can be an effective and flexible solution to generating a large number of items, with few constraints on item type, language, subject domain, or the data source used for training LLMs to create items. However, we did not find many studies reporting the measurement properties of the generated items.

4. DISCUSSION and CONCLUSION

4.1. The Current State of Research on AIG

As a summary of the findings from this review, we identified the most commonly used LLMs in the current AIG literature, such as T5, BERT, GPT, and their variants. We described the characteristics and features of each base type of LLM, linking them to their specific uses in the AIG process. LLMs used in the pre-generation stage focus on preparing, processing, and understanding texts for subsequent item generation to ensure high quality. In the post-generation stage, LLMs are primarily used to filter out low-quality items (i.e., mostly focusing on the correctness of grammar and syntax as well as the semantic relevancy and similarity) or determine the usefulness of the generated items.

After reviewing the existing studies, we conclude that LLMs prove useful in generating large banks of items. Additionally, we revealed that LLMs offer a highly flexible solution for AIG, as they have virtually no constraints in terms of item type, language, the subject domain of the items to be generated, or the data source used for further training of LLMs.

4.2. Current Research Gaps in the Literature

While the reviewed studies often show that LLMs are effective and flexible in creating a large number of items, we found that many studies applying LLMs in AIG often lacked a solid educational foundation. This might be because many of the authors were NLP researchers who possessed limited recognition and knowledge of learning or measurement theories. Alternatively, it could be because creating high-quality items that are readily usable for educational contexts was not their primary interest or research focus. Accordingly, many of those items are generated without deep consideration of their measurement purposes and item properties, which are essential to meet the requirements of educational assessment.

In traditional item development or template-based AIG, item generation starts with a clear definition of what to measure (i.e., identifying the construct to be measured by considering the expected learner outcomes and instructional objectives), why to measure (i.e., the purpose of assessment), and how to measure (i.e., assessment design and item format), while only a few studies leveraging LLMs for AIG considered these important aspects. For example, many of those generated items in the existing studies do not attempt to evaluate the higher-level cognitive processes specified in Bloom's taxonomy, such as applying, analyzing, evaluating, or creating. This is because they are mostly created by researchers from the AIG area whose original purpose was to develop chatbots and conversational agents. Therefore, their item generation predominantly focuses on the levels of remembering or understanding, similar to the goal of conversational agents, which does not always meet or fit the measurement purposes and goals that are to evaluate the complex learning progress and outcomes of human students.

Moreover, while some studies invited human participants to evaluate the quality of the items after being generated, only a few involved SMEs or measurement specialists in the AIG process. However, human experts are crucial for guiding item development, ensuring quality, and potentially enhancing students' learning outcomes through formative use (e.g., Gierl & Lai, 2012; Lu *et al.*, 2021). The absence of expert guidance has led many existing AIG studies using LLMs to overlook important measurement properties such as item difficulty and item discrimination. Thus, the current literature mostly provides evidence that LLMs can be leveraged to generate a large number of items, but little is known about whether these items possess the high quality necessary for educational purposes such as pedagogical teaching and

assessment. Considering the goal of AIG, which is to generate large banks of high-quality items while reducing overall costs (Gierl & Lai, 2012; Lai *et al.*, 2009), we argue that a thorough item evaluation after generation is missing in the current literature. That is, AIG does not end up generating a large number of items but ensuring that these items are of high quality and can fulfill their demands for educational purposes and contexts. Thus, we argue that AIG is still a developing and promising research domain, and its further development and application will depend on the involvement of human experts and the integration of learning and measurement theories.

4.3. Suggestions for Future Research

The current research gaps, particularly the lack of educational foundations in the reviewed AIG studies, led us to advocate for two suggestions for future research in this area.

First, future research should prioritize clarifying the assessment context and measurement goals in AIG applications. While AIG benefits from advancements in NLP, such as LLMs, it differs from general text generation by creating items that meet the specific purposes of assessments in real educational contexts. Newton (2007) distinguished 18 different educational assessment applications, such as formative, diagnostic, qualification certification, and comparative purposes. For example, formative assessments take place during students' learning processes and support learning by providing effective feedback, while summative assessments aim to collect information about students' learning outcomes after the learning process has occurred. As the assessment purposes differ, the desired characteristics of the generated items also differ. When generating items for a specific educational and assessment context, researchers must identify the nature and desired characteristics of these items. For instance, in the case of formative assessments aimed at aiding students' learning, items should prioritize pedagogical value by providing feedback to help identify misconceptions. To assess students' learning outcomes following a teaching program, items should be balanced and cover a broad range of key concepts taught, rather than disproportionately focusing on concepts within a narrow content area. Furthermore, regardless of whether items are formative or summative or created for other assessment purposes, they must demonstrate high quality, with strong reliability and validity, to be effective, because they directly impact the feedback provided to students or the ability to draw inferences about students' learning progress and outcomes.

Second, we recommend evaluating both the measurement properties and pedagogical soundness of generated items as an essential step in AIG. From a practical standpoint, item development does not end with item generation; the evaluation of item quality is crucial to ensure usability in educational contexts. For example, measurement properties of items (e.g., difficulty and discrimination parameters) and tests (e.g., reliability and validity) can be examined using measurement theories such as classical test theory and item response theory. Finally, after items are generated and evaluated for their measurement properties, the focus should shift back to meeting the intended measurement purposes or pedagogical value of assessments. Educators and SMEs can assess the educational or pedagogical value of automatically generated items. It is important to ensure that the generated items can effectively serve their intended purpose. If not, it is necessary to revisit previous item development stages for revisions to create items that better align with the learning objectives and measurement goals.

The three-stage AIG framework emerging from our analysis (i.e., pre-generation stage, item generation stage, post-generation stage) offers a structured approach for integrating these two recommendations into practice. This framework not only categorizes LLM applications but also has a potential for seamlessly integrating human expertise for a human-in-the-loop approach to strengthen educational foundations of AIG practices. In the pre-generation stage, which focuses on preparation tasks before item generation, AIG practitioners can enhance outcomes by incorporating human expertise. For example, they may define assessment goals and formats or

contribute to model fine-tuning by providing context-relevant educational materials (e.g., item banks). Such practices have been shown to improve both the model's contextual awareness and task-specific performance (Ghanem *et al.*, 2022; Ratcheva *et al.*, 2022). During the item generation stage, human experts (e.g., educators) can refine prompts to elicit more appropriate items aligned with specific learning or assessment goals. Finally, in the post-generation stage, subject-matter experts, students (as target test-takers), and even LLMs themselves may participate in evaluating the generated items.

While not explicitly framed in these terms, Sayin and Gierl's (2024) study demonstrates strong alignment with this three-stage approach when analyzed through our conceptual framework. Their methodology began with what we would characterize as pre-generation preparation, involving the development of three foundational models: a cognitive model to define target knowledge and skills, an item model to establish construction guidelines, and a specialized text analysis model to ensure passage coherence and plausible distractors. These models then informed precise prompt engineering for GPT-3.5, enabling systematic generation of 12,500 grade-level items for high-stakes examinations - a process corresponding to our item generation stage. Finally, their comprehensive quality assurance procedures, including expert reviews and empirical field testing to assess critical psychometric properties like item difficulty and discrimination indices, exemplify what our framework identifies as post-generation validation.

4.4. Limitations

We did not conduct a formal assessment of the methodological quality of the included studies, which is generally considered a limitation of scoping reviews. This is because the primary purpose of this study, and of scoping reviews in general, is to map the existing literature on a particular topic to explore the range, nature, and extent of research activities related to the topic, rather than to evaluate the quality of the studies (Arksey & O'Malley, 2005). Additionally, we acknowledge that the field of leveraging LLMs is rapidly evolving. As we completed our keyword search, new LLMs have been introduced, boasting significantly larger parameter sizes and enhanced functionalities. This rapid development has not only expanded the utility and potential of LLMs but has also made advanced technology more accessible to users with varying levels of programming skills. Consequently, the use of LLMs in education continues to evolve, and understanding how these models can best serve the AIG process remains an ongoing journey.

4.5. Conclusion

As a category of generative artificial intelligence technology, LLMs focus on language understanding and text generation, which has sparked researchers' interest in using them to automatically generate questions for educational purposes. Through the mapping and synthesis of the reviewed studies on leveraging LLMs for AIG, we identified that the most commonly used LLMs are T5, BERT, GPT, and their variants. We have also categorized the current applications of LLMs into three stages of the AIG process: pre-generation, item generation, and post-generation. The findings reveal that using LLMs to generate items is an effective and flexible solution, with few constraints on item type, language, subject domain, or the data source used for training LLMs to create items. Due to the exceptional language understanding abilities of LLMs, the generated items are typically free from grammar errors and contextually relevant to the desired content domain. However, we also noted a lack of a solid educational foundation in many of the reviewed studies, as they did not incorporate learning and measurement theories into the item generation process. We attribute this issue to the absence of involvement of human experts such as SMEs and measurement specialists. Although one part of the goal of AIG was to reduce financial costs and human burdens, it was never meant to exclude human involvement from the item development process. Importantly, AIG is still considered an augmented intelligence approach, which means it requires both human expertise and the capabilities of a computer. Considering the goal of AIG, we not only want to generate a large number of items

but also care more about the item quality and characteristics. Hence, future researchers should consider enhancing the educational foundation in the AIG process. The three-stage framework proposed in this review provides a structured approach to integrating human expertise at each stage—pre-generation, item generation, and post-generation—to ensure high-quality item generation. This way, LLMs can be leveraged to produce items that are not only grammatically correct and contextually relevant but also reliable, valid, and pedagogically sound.

Acknowledgments

This work has been previously presented at the 2024 National Council on Measurement in Education (NCME) Annual Meeting, Philadelphia, Pennsylvania.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors.

Contribution of Authors

Bin Tan: Conceptualization, methodology, literature search, article screening, formal analysis and investigation, writing – original draft preparation, writing – review and editing. **Nour Armoush:** Conceptualization, methodology, literature search, article screening, formal analysis and investigation, writing – original draft preparation. **Elisabetta Mazzullo:** Formal analysis and investigation, writing – original draft preparation. **Okan Bulut:** Conceptualization, methodology, writing – review and editing, supervision. **Mark Gierl:** Conceptualization, methodology, writing – review and editing, supervision.

Orcid

Bin Tan  <https://orcid.org/0000-0001-6717-5620>

Nour Armoush  <https://orcid.org/0009-0008-2310-5098>

Elisabetta Mazzullo  <https://orcid.org/0009-0008-4847-9934>

Okan Bulut  <https://orcid.org/0000-0001-5853-1267>

Mark J. Gierl  <https://orcid.org/0000-0001-6717-5620>

REFERENCES

- Ackerman, R., & Balyan, R. (2023). Automatic multilingual question generation for health data using LLMs. In *International Conference on AI-generated Content* (pp. 1-11). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-99-7587-7_1
- Agrawal, A., & Shukla, P. (2023). Context aware automatic subjective and objective question generation using Fast Text to text transfer learning. *International Journal of Advanced Computer Science and Applications*, 14(4), 456-463.
- Aigo, K., Tsunakawa, T., Nishida, M., & Nishimura, M. (2021). Question generation using knowledge graphs with the T5 language model and masked self-attention. In *2021 IEEE 10th Global Conference on Consumer Electronics* (pp. 85-87). IEEE. <https://doi.org/10.1109/GCCE53005.2021.9621874>
- Akyön, F.Ç., Cavusoglu, A.D.E., Cengiz, C., Altinuç, S.O., & Temizel, A. (2022). Automated question generation and question answering from Turkish texts. *Turkish Journal of Electrical Engineering and Computer Sciences*, 30(5), 1931-1940. <https://doi.org/10.55730/1300-0632.3914>
- Alsubait, T., Parsia, B., & Sattler, U. (2016). Ontology-based multiple choice question generation. *KI-Künstliche Intelligenz*, 30, 183-188. <https://doi.org/10.1007/s13218-015-0405-9>
- Alves, C.B., Gierl, M.J., & Lai, H. (2010, April). *Using automated item generation to promote test design and development* [Paper presentation]. American Educational Research Association Annual Meeting, Denver, CO, United States.

- Arksey, H., & O'malley, L. (2005). Scoping studies: towards a methodological framework. *International Journal of Social Research Methodology*, 8(1), 19-32. <https://doi.org/10.1080/1364557032000119616>
- Attali, Y., Runge, A., LaFlair, G.T., Yancey, K., Goodwin, S., Park, Y., & Von Davier, A.A. (2022). The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence*, 5, 903077. <https://doi.org/10.3389/frai.2022.903077>
- Berger, G., Rischewski, T., Chiruzzo, L., & Rosá, A. (2022). Generation of English question answer exercises from texts using transformers-based models. In *2022 IEEE Latin American Conference on Computational Intelligence* (pp. 1-5). IEEE. <https://doi.org/10.1109/LA-CCI54402.2022.9981171>
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1), 7-74. <https://doi.org/10.1080/0969595980050102>
- Bulathwela, S., Muse, H., & Yilmaz, E. (2023). Scalable educational question generation with pre-trained language models. In *International Conference on Artificial Intelligence in Education* (pp. 327-339). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-36272-9_27
- Bulut, O., & Yildirim-Erbasli, S.N. (2022). Automatic story and item generation for reading comprehension assessments with transformers. *International Journal of Assessment Tools in Education*, 9(Special Issue), 72-87. <https://doi.org/10.21449/ijate.1124382>
- Bulut, O., Gorgun, G., Yildirim-Erbasli, S.N., Wongvorachan, T., Daniels, L.M., Gao, Y., ... & Shin, J. (2023). Standing on the shoulders of giants: Online formative assessments as the foundation for predictive learning analytics models. *British Journal of Educational Technology*, 54(1), 19-39. <https://doi.org/10.1111/bjet.13276>
- Ch, D.R., & Saha, S.K. (2018). Automatic multiple choice question generation from text: A survey. *IEEE Transactions on Learning Technologies*, 13(1), 14-25. <https://doi.org/10.1109/TLT.2018.2889100>
- Chiang, S.H., Wang, S.C., & Fan, Y.C. (2024). Cdgp: Automatic cloze distractor generation based on pre-trained language model. *arXiv preprint arXiv:2403.10326*. <https://doi.org/10.18653/v1/2022.findings-emnlp.429>
- Chughtai, R., Azam, F., Anwar, M.W., But, W.H., & Farooq, M.U. (2022). A lecture centric automated distractor generation for post-graduate software engineering courses. In *2022 International Conference on Frontiers of Information Technology (FIT)* (pp. 100-105). IEEE. <https://doi.org/10.1109/FIT57066.2022.00028>
- Chung, H.L., Chan, Y.H., & Fan, Y.C. (2020). A BERT-based distractor generation scheme with multi-tasking and negative answer training strategies. *arXiv preprint arXiv:2010.05384*. <https://arxiv.org/abs/2010.05384>
- Dalby, D., & Swan, M. (2019). Using digital technology to enhance formative assessment in mathematics classrooms. *British Journal of Educational Technology*, 50(2), 832-845. <https://doi.org/10.1111/bjet.12606>
- Dembitzer, L., Zelikovitz, S., & Kettler, R.J. (2017). Designing computer-based assessments: Multidisciplinary findings and student perspectives. *International Journal of Educational Technology*, 4(3), 20-31. <https://educationaltechnology.net/ijet/index.php/ijet/article/view/47>
- Desai, T. (2021). *Discourse parsing and its application to question generation* [Unpublished dissertation]. The University of Texas at Dallas.
- Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. <https://doi.org/10.48550/arXiv.1810.04805>
- Dijkstra, R., Genç, Z., Kayal, S., & Kamps, J. (2022). Reading comprehension quiz generation using generative pre-trained transformers. In S. Sosnovsky, P. Brusilovsky, & A. Lan (Eds.), *Proceedings of the Fourth International Workshop on Intelligent Textbooks 2022* (pp. 4–7). CEUR-WS. <http://ceur-ws.org/Vol-3192/>

- Drori, I., Zhang, S., Shuttleworth, R., Tang, L., Lu, A., Ke, E., ... & Strang, G. (2022). A neural network solves, explains, and generates university math problems by program synthesis and few-shot learning at human level. *Proceedings of the National Academy of Sciences*, 119(32), e2123433119. <https://doi.org/10.1073/pnas.2123433119>
- Falcão, F., Costa, P., & Pêgo, J.M. (2022). Feasibility assurance: a review of automatic item generation in medical assessment. *Advances in Health Sciences Education*, 27(2), 405-425. <https://doi.org/10.1007/s10459-022-10092-z>
- Femi, J.G., & Nayak, A.K. (2022). EQGTL: An Ensemble Model for Relevant Question Generation using Transfer Learning. In *2022 International Conference on Machine Learning, Computer Systems and Security* (pp. 253-258). IEEE. <https://doi.org/10.1109/MLCSS57186.2022.00054>
- Fuadi, M., & Wibawa, A.D. (2022). Automatic question generation from indonesian texts using text-to-text transformers. In *2022 International Conference on Electrical and Information Technology (IEIT)* (pp. 84-89). IEEE. <https://doi.org/10.1109/IEIT56384.2022.9967858>
- Fung, Y.C., Kwok, J.C.W., Lee, L.K., Chui, K.T., & U, L.H. (2020). Automatic question generation system for english reading comprehension. In *Technology in Education. Innovations for Online Teaching and Learning: 5th International Conference, ICTE 2020, Macau, China, August 19-22, 2020, Revised Selected Papers 5* (pp. 136-146). Springer Singapore. https://doi.org/10.1007/978-981-33-4594-2_12
- Fung, Y.C., Lee, L.K., & Chui, K.T. (2023). An automatic question generator for Chinese comprehension. *Inventions*, 8(1), 31. <https://doi.org/10.3390/inventions8010031>
- Ghanem, B., Coleman, L.L., Dexter, J.R., von der Ohe, S.M., & Fyshe, A. (2022). Question generation for reading comprehension assessment by modeling how and what to ask. *arXiv preprint arXiv:2204.02908*. <https://doi.org/10.48550/arXiv.2204.02908>
- Gierl, M.J., & Lai, H. (2012). The role of item models in automatic item generation. *International Journal of Testing*, 12(3), 273-298. <https://doi.org/10.1080/15305058.2011.635830>
- Gierl, M.J., & Lai, H. (2015). Automatic item generation. In *Handbook of test development* (pp. 410-429). Routledge.
- Gierl, M.J., & Lai, H. (2016). A process for reviewing and evaluating generated test items. *Educational Measurement: Issues and Practice*, 35(4), 6-20. <https://doi.org/10.1111/emip.12129>
- Gierl, M.J., Lai, H., & Tanygin, V. (2021). *Advanced methods in automatic item generation*. Routledge.
- Godslove, J.F., & Nayak, A.K. (2023). Generative model for formulating relevant questions and answers using transfer learning. In *AIP Conference Proceedings* (Vol. 2819, No. 1). AIP Publishing. <https://doi.org/10.1063/5.0136892>
- Gopal, A. (2022). Automatic question generation for Hindi and Marathi. In *2022 International Conference on Advanced Learning Technologies (ICALT)* (pp. 19-21). IEEE. <https://doi.org/10.1109/ICALT55010.2022.00012>
- Goyal, R., Kumar, P., & Singh, V.P. (2023). Automated question and answer generation from texts using text-to-text transformers. *Arabian Journal for Science and Engineering*, 1-15. <https://doi.org/10.1007/s13369-023-07840-7>
- Granić, A. (2022). Educational technology adoption: A systematic review. *Education and Information Technologies*, 27(7), 9725-9744. <https://doi.org/10.1007/s10639-022-10951-7>
- Grover, K., Kaur, K., Tiwari, K., Rupali, & Kumar, P. (2021). Deep learning based question generation using t5 transformer. In *Advanced Computing: 10th International Conference, IACC 2020, Panaji, Goa, India, December 5–6, 2020, Revised Selected Papers, Part I 10* (pp. 243-255). Springer Singapore. https://doi.org/10.1007/978-981-16-0401-0_18
- Han, Z. (2022). *Unsupervised multilingual distractor generation for fill-in-the-blank questions* [Unpublished thesis]. Uppsala University.

- Jiao, Y., Shridhar, K., Cui, P., Zhou, W., & Sachan, M. (2023). Automatic educational question generation with difficulty level controls. In *International Conference on Artificial Intelligence in Education* (pp. 476-488). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-36272-9_39
- Kalpakchi, D., & Boye, J. (2021). BERT-based distractor generation for Swedish reading comprehension questions using a small-scale dataset. *arXiv preprint arXiv:2108.03973*. <https://doi.org/10.48550/arXiv.2108.03973>
- Kasakowskij, R., Kasakowskij, T., & Seidel, N. (2022). Generation of multiple true false questions. 20. Fachtagung Bildungstechnologien. <https://doi.org/10.18420/delfi2022-026>
- Khandait, K., Bhura, S., & Asole, S.S. (2022). Automatic question generation through word vector synchronization using llama. *Indian Journal of Computer Science and Engineering*, 13(4), 1083-1095. <https://doi.org/10.21817/indjcse/2022/v13i4/221304046>
- Kosh, A.E., Simpson, M.A., Bickel, L., Kellogg, M., & Sanford-Moore, E. (2019). A cost–benefit analysis of automatic item generation. *Educational Measurement: Issues and Practice*, 38(1), 48-53. <https://doi.org/10.1111/emip.12237>
- Kumar, A., Kharadi, A., Singh, D., & Kumari, M. (2021). Automatic question-answer pair generation using deep learning. In *2021 Third International Conference on Inventive Research in Computing Applications* (pp. 794-799). IEEE. <https://doi.org/10.1109/ICIRCA51532.2021.9544654>
- Kumar, N.S., Mali, R., Ratnam, A., Kurpad, V., & Magapu, H. (2022). Identification and addressal of knowledge gaps in students. In *2022 3rd International Conference for Emerging Technology* (pp. 1-6). IEEE. <https://doi.org/10.1109/INCET54531.2022.9824483>
- Kumar, S., Chauhan, A., & Kumar C, P. (2022). Learning enhancement using question-answer generation for e-book using contrastive fine-tuned T5. In *International Conference on Big Data Analytics* (pp. 68-87). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-24094-2_5
- Kumari, V., Keshari, S., Sharma, Y., & Goel, L. (2022). Context-based question answering system with suggested questions. In *2022 12th International Conference on Cloud Computing, Data Science & Engineering* (pp. 368-373). IEEE. <https://doi.org/10.1109/Confluence52989.2022.9734207>
- Kuo, C.Y., & Wu, H.K. (2013). Toward an integrated model for designing assessment systems: An analysis of the current status of computer-based assessments in science. *Computers & Education*, 68, 388-403. <https://doi.org/10.1016/j.compedu.2013.06.002>
- Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2020). A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30, 121-204. <https://doi.org/10.1007/s40593-019-00186-y>
- Lai, H., Alves, C., & Gierl, M.J. (2009). Using automatic item generation to address item demands for CAT. In D.J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*. www.psych.umn.edu/psylabs/CATCentral/
- Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*. <https://doi.org/10.48550/arXiv.1909.11942>
- Lee, H., Chung, H.Q., Zhang, Y., Abedi, J., & Warschauer, M. (2020). The effectiveness and features of formative assessment in US K-12 education: A systematic review. *Applied Measurement in Education*, 33(2), 124-140. <https://doi.org/10.1080/08957347.2020.1732383>
- Lim, Y.S. (2019). Students' perception of formative assessment as an instructional tool in medical education. *Medical Science Educator*, 29(1), 255-263. <https://doi.org/10.1007/s40670-018-00687-w>
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*. <https://doi.org/10.48550/arXiv.1907.11692>

- Lu, O.H.T., Huang, A.Y.Q., Tsai, D.C.L., & Yang, S.J.H. (2021). Expert-authored and machine-generated short- answer questions for assessing students' learning performance. *Educational Technology & Society*, 24(3), 159–173. <https://www.jstor.org/stable/27032863>
- Maheen, F., Asif, M., Ahmad, H., Ahmad, S., Alturise, F., Asiry, O., & Ghadi, Y.Y. (2022). Automatic computer science domain multiple-choice questions generation based on informative sentences. *PeerJ Computer Science*, 8, e1010. <https://doi.org/10.7717/peerj-cs.1010>
- Malhar, A., Sawant, P., Chhadva, Y., & Kurhade, S. (2022). Deep learning-based Answering Questions using T5 and Structured Question Generation System. In *2022 6th International Conference on Intelligent Computing and Control Systems* (pp. 1544-1549). IEEE. <https://doi.org/10.1109/ICICCS53718.2022.9788264>
- Mathur, A., & Suchithra, M. (2022). Application of abstractive summarization in multiple choice question generation. In *2022 International Conference on Computational Intelligence and Sustainable Engineering Solutions* (pp. 409-413). IEEE. <https://doi.org/10.1109/CISE554857.2022.9844396>
- Matsumori, S., Okuoka, K., Shibata, R., Inoue, M., Fukuchi, Y., & Imai, M. (2023). Mask and cloze: Automatic open cloze question generation using a masked language model. *IEEE Access*, 11, 9835-9850. <https://doi.org/10.1109/ACCESS.2023.3239005>
- Maurya, K.K., & Desarkar, M.S. (2020). Learning to distract: A hierarchical multi-decoder network for automated generation of long distractors for multiple-choice questions for reading comprehension. In *Proceedings of the 29th ACM international conference on information & knowledge management* (pp. 1115-1124). <https://doi.org/10.1145/3340531.3411997>
- Mazzullo, E., Bulut, O., Wongvorachan, T., & Tan, B. (2023). Learning Analytics in the Era of Large Language Models. *Analytics*, 2(4), 877-898. <https://doi.org/10.3390/analytics2040046>
- Min, B., Ross, H., Sulem, E., Veyseh, A.P.B., Nguyen, T.H., Sainz, O., ... & Roth, D. (2023). Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2), 1-40. <https://doi.org/10.1145/3605943>
- Muse, H., Bulathwela, S., & Yilmaz, E. (2022). Pre-training with scientific text improves educational question generation. *arXiv preprint arXiv:2212.03869*. <https://doi.org/10.48550/arXiv.2212.03869>
- Newton, P.E. (2007). Clarifying the purposes of educational assessment. *Assessment in education*, 14(2), 149-170. <https://doi.org/10.1080/09695940701478321>
- Nguyen, H.A., Bhat, S., Moore, S., Bier, N., & Stamper, J. (2022). Towards generalized methods for automatic question generation in educational domains. In *European conference on technology enhanced learning* (pp. 272-284). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-16290-9_20
- Nittala, S., Agarwal, P., Vishnu, R., & Shanbhag, S. (2022). Speaker Diarization and BERT-Based Model for Question Set Generation from Video Lectures. In *Information and Communication Technology for Competitive Strategies ICT: Applications and Social Interfaces* (pp. 441-452). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-19-0095-2_42
- Offerijns, J., Verberne, S., & Verhoef, T. (2020). Better distractions: Transformer-based distractor generation and multiple-choice question filtering. *arXiv preprint arXiv:2010.09598*. <https://doi.org/10.48550/arXiv.2010.09598>
- Pochiraju, D., Chakilam, A., Betham, P., Chimulla, P., & Rao, S.G. (2023). Extractive summarization and multiple-choice question generation using XLNet. In *2023 7th International Conference on Intelligent Computing and Control Systems* (pp. 1001-1005). IEEE. <https://doi.org/10.1109/ICICCS56967.2023.10142220>
- Pugh, D., De Champlain, A., Gierl, M., Lai, H., & Touchie, C. (2020). Can automated item generation be used to develop high quality MCQs that assess application of knowledge?.

- Research and Practice in Technology Enhanced Learning*, 15, 1-13. <https://doi.org/10.1186/s41039-020-00134-8>
- Qiu, X., Xue, H., Liang, L., Xie, Z., Liao, S., & Shi, G. (2021). Automatic generation of multiple-choice cloze-test questions for lao language learning. In *2021 International Conference on Asian Language Processing* (pp. 125-130). IEEE. <https://doi.org/10.1109/IALP54817.2021.9675153>
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140), 1-67. <https://doi.org/10.48550/arXiv.1910.10683>
- Raina, V., & Gales, M. (2022). Multiple-choice question generation: Towards an automated assessment framework. *arXiv preprint arXiv:2209.11830*. <https://doi.org/10.48550/arXiv.2209.11830>
- Ratcheva, M.G., Navale, R., & Desai, B.C. (2022). An online MCQ sub-system for CrsMgr. In *Proceedings of the 26th International Database Engineered Applications Symposium* (pp. 128-133). <https://doi.org/10.1145/3548785.3548789>
- Rodriguez-Torrealba, R., Garcia-Lopez, E., & Garcia-Cabot, A. (2022). End-to-end generation of multiple-choice questions using text-to-text transfer transformer models. *Expert Systems with Applications*, 208, 118258. <https://doi.org/10.1016/j.eswa.2022.118258>
- Rudner, L.M. (2009). Implementing the graduate management admission test computerized adaptive test. In *Elements of adaptive testing* (pp. 151-165). Springer New York. https://doi.org/10.1007/978-0-387-85461-8_8
- Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*. <https://doi.org/10.48550/arXiv.1910.01108>
- Sayin, A., & Gierl, M. (2024). Using OpenAI GPT to generate reading comprehension items. *Educational Measurement: Issues and Practice*, 43(1), 5-18. <https://doi.org/10.1111/emip.12590>
- Shan, J., Nishihara, Y., Maeda, A., & Yamanishi, R. (2022). Question generation for reading comprehension test complying with types of question. *Journal of Information Science & Engineering*, 38(3). [https://doi.org/10.6688/JISE.202205_38\(3\).0005](https://doi.org/10.6688/JISE.202205_38(3).0005)
- Shan, J., Nishihara, Y., Yamanishi, R., & Maeda, A. (2019). Question generation for reading comprehension of language learning test: A method using Seq2Seq approach with transformer model. In *2019 International Conference on Technologies and Applications of Artificial Intelligence* (pp. 1-6). IEEE. <https://doi.org/10.1109/TAAI48200.2019.8959903>
- Shridhar, K., Macina, J., El-Assady, M., Sinha, T., Kapur, M., & Sachan, M. (2022). Automatic generation of socratic subquestions for teaching math word problems. *arXiv preprint arXiv:2211.12835*. <https://doi.org/10.48550/arXiv.2211.12835>
- Singh, J., McCann, B., Socher, R., & Xiong, C. (2019). BERT is not an interlingua and the bias of tokenization. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)* (pp. 47-55). <https://doi.org/10.18653/v1/D19-6106>
- Spector, J.M., Ifenthaler, D., Samspon, D., Yang, L., Mukama, E., Warusavitarana, A., ... Gibson, D.C. (2016). Technology enhanced formative assessment for 21st century learning. *Educational Technology & Society*, 19(3), 58-71. <https://www.jstor.org/stable/jeductechsoc.i.19.3.58>
- Srivastava, M., & Goodman, N. (2021). Question generation for adaptive education. *arXiv preprint arXiv:2106.04262*. <https://doi.org/10.48550/arXiv.2106.04262>
- Steuer, T., Filighera, A., & Rensing, C. (2020). Exploring artificial jabbering for automatic text comprehension question generation. In *Addressing Global Challenges and Quality Education: 15th European Conference on Technology Enhanced Learning, EC-TEL 2020, Heidelberg, Germany, September 14–18, 2020, Proceedings 15* (pp. 1-14). Springer International Publishing. https://doi.org/10.1007/978-3-030-57717-9_1

- Tsai, D.C., Chang, W., & Yang, S. (2021). Short answer questions generation by Fine-Tuning BERT and GPT-2. In *Proceedings of the 29th International Conference on Computers in Education Conference* (Vol. 64). <https://icce2021.apsce.net/wp-content/uploads/2021/12/ICCE2021-Vol.II-PP.-508-514.pdf>
- von Davier, M. (2019). Training Optimus prime, MD: Generating medical certification items by fine-tuning OpenAI's gpt2 transformer model. *arXiv preprint arXiv:1908.08594*. <https://doi.org/10.48550/arXiv.1908.08594>
- Vu, N., & Van Nguyen, K. (2022). Enhancing Vietnamese question generation with reinforcement learning. In *Asian Conference on Intelligent Information and Database Systems* (pp. 559-570). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-21743-2_45
- Wang, B., Yao, T., Chen, W., Xu, J., & Wang, X. (2021). Multi-lingual question generation with language agnostic language model. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (pp. 2262-2272). <https://aclanthology.org/2021.findings-acl.199.pdf>
- Wang, H.C., Maslim, M., & Kan, C.H. (2023). A question–answer generation system for an asynchronous distance learning platform. *Education and Information Technologies*, 28(9), 12059-12088. <https://doi.org/10.1007/s10639-023-11675-y>
- Wang, Z., Valdez, J., Basu Mallick, D., & Baraniuk, R.G. (2022). Towards human-like educational question generation with large language models. In *International conference on artificial intelligence in education* (pp. 153-166). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-11644-5_13
- Wylie, E.C., & Lyon, C.J. (2015). The fidelity of formative assessment implementation: Issues of breadth and quality. *Assessment in Education: Principles, Policy & Practice*, 22(1), 140-160. <https://doi.org/10.1080/0969594X.2014.990416>
- Xie, J., Peng, N., Cai, Y., Wang, T., & Huang, Q. (2021). Diverse distractor generation for constructing high-quality multiple choice questions. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30, 280-291. <https://doi.org/10.1109/TASLP.2021.3138706>
- Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., & Le, Q.V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32. <https://dl.acm.org/doi/10.5555/3454287.3454804>
- Yen, Y.-C., Ho, R.-G., Liao, W.-W., & Chen, L.-J. (2012). Reducing the impact of inappropriate items on reviewable computerized adaptive testing. *Educational Technology & Society*, 15(2), 231–243. <https://www.jstor.org/stable/jeductechsoci.15.2.231>
- Zhang, C. (2023). Automatic Generation of Multiple-Choice Questions. *arXiv preprint arXiv:2303.14576v1*. <https://doi.org/10.48550/arXiv.2303.14576>
- Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J.R. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*. <https://doi.org/10.48550/arXiv.2303.18223>
- Zhao, Z., Hou, Y., Wang, D., Yu, M., Liu, C., & Ma, X. (2022). Educational question generation of children storybooks via question type distribution learning and event-centric summarization. *arXiv preprint arXiv:2203.14187*. <https://doi.org/10.48550/arXiv.2203.14187>

APPENDIX A

Summary of the included studies

Authors	Venue	Description of the use of LLMs
Agrawal & Shukla, 2023	International Journal of Advanced Computer Science and Applications	T5 was trained to generate questions using texts and answers together as input.
Aigo et al., 2021	Global Conference on Consumer Electronics	T5 was used for question generation based on knowledge graphs.
Akyön et al., 2022	Turkish Journal of Electrical Engineering and Computer Sciences	mT5 was trained to extract answers first, then use these as input for generating questions. Other models were used for comparisons.
Attali et al., 2022	Frontiers in Artificial Intelligence	GPT-3 was first used to generate question generation instructions, which were then used to generate associated questions and answers.
Berger et al., 2022	Latin American Conference on Computational Intelligence	T5-small was used to generate questions with text-answer pairs as input.
Bulathwela et al., 2023	arxiv	T5 was used to generate questions with raw question-answer pair texts as input
Bulut & Yildirim-Erbasli, 2022	International Journal of Assessment Tools in Education	GPT was used to predict the next word for generating reading stories with two prompts, while T5 was utilized for text-to-text generation with texts as input to generate questions.
Chiang et al., 2022	Empirical Methods in Natural Language Processing (Conference)	BERT was used with a QA pair as input and a distractor candidate as output, which were then filtered by a text embedding model.
Chughtai et al., 2022	International Conference on Frontiers of Information Technology	T5 was used to tokenize and summarize texts, and to generate question-answer pairs using contexts and paragraphs as input.
Chung et al., 2020	arxiv	BERT was employed to iteratively predict the next token of distractors based on the context, question, correct answer, and previously predicted distractor tokens. GPT was used as a baseline model.
Desai, 2021	Doctoral dissertation	BERT was used for discourse segmentation, textual representation, and discourse parsing (identifying connections of tokens). However, the question generation process still relies on template-based methods with masked tokens/words.
Dijkstra et al., 2022	International Workshop on Intelligent Textbooks	Two prompts were compared for GPT to generate questions: (1) Questions were generated based on the context first and then combined with the generated answers to create distractors. (2) GPT directly generated the quiz based on the context alone.
Drori et al., 2022	arxiv	The LLMs were used to generate questions using two prompting strategies: one shot vs. few shots examples.
Femi & Nayak, 2022	International Conference on Machine Learning, Computer Systems and Security	BERT was initially used to make sense of the keywords generated by WordNet and to put the keywords in context, followed by T5 generating questions, answers, and distractors.
Fuadi & Wibawa, 2022	International Conference on Electrical and Information Technology	T5 was firstly used to generate answers based on contexts, and these answers, along with the contexts, were then used to generate questions.

Fung <i>et al.</i> , 2020	International Conference on Technology in Education	T5 was trained to generate questions using a source sentence with an answer phrase as input.
Fung <i>et al.</i> , 2023	Inventions	T5 was used for question generation with contexts and answers as input. The generated questions were compared with existing questions for feedback, which fed into the next iteration of question generation.
Ghanem <i>et al.</i> , 2022	arxiv	T5 was trained to teach the model how to ask questions (a skill label), then used to generate questions with a story as input and QA pairs as output.
Godslove & Nayak, 2023	International Conference on Applied Mathematics in Science and Engineering	BertWSD was used for word sense disambiguation, while T5 was utilized to generate questions, answers, and distractors with processed texts as input (including text summarization and extraction).
Gopal, 2022	International Conference on Advanced Learning Technologies	T5 and GPT-2 were used to generate texts with input texts.
Goyal <i>et al.</i> , 2023	Arabian Journal for Science and Engineering	T5 was used to identify answers and was trained to generate answers with texts as input, which were then combined with these answers to generate questions.
Grover <i>et al.</i> , 2021	International Advanced Computing Conference	T5 was trained to generate multiple questions by providing context paragraphs.
Han, 2022	Master's thesis	BERT variants and BART variants were used to generate foreign language distractors following the masked token prediction method. The LLMs were also used to generate distractors.
Jiao <i>et al.</i> , 2023	International Conference on Artificial Intelligence in Education	T5, BERT, and GPT were compared for AIG tasks. In addition, GPT-2 was used to calculate perplexity values, indicating the fluency of the generated items. BERT-large was also used to evaluate coherence. BertScore was employed to calculate semantic differences in order to assess creativity.
Kalpakchi & Boye, 2021	arxiv	BERT was used to generate questions with question, context, and correct answers as input in two modes: sequential and non-ordered.
Kasakowskij <i>et al.</i> , 2022	Fachtagung Bildungstechnologien (DELFI)	GPT-2 was used to generate false statements for true/false questions based on the extracted correct statements from texts.
Khandait <i>et al.</i> , 2022	Indian Journal of Computer Science and Engineering	T5 models received input paragraphs to generate questions. Other LLMs were used as baseline models without detailed information for implementation.
Kumar <i>et al.</i> , 2021	International Conference on Inventive Research in Computing Applications	First, T5 was employed to extract answers from given passages. Next, T5 was used to generate questions with passages and answers as inputs.
Kumar <i>et al.</i> , 2022	International Conference for Emerging Technology	T5 was used to summarize texts and to generate single-line questions with the candidate sentences as input and QA pairs as outputs. GPT-2 was used to generate true/false questions. Alternative sentences were firstly generated by GPT-2, followed by using Sentence BERT to filter out alternative sentences that are too similar to the candidate sentences. Sentence BERT was also used to calculate the similarity scores of distractors for question filtering.
Kumar <i>et al.</i> ,	International Conference on	T5 was used to summarize the raw texts and generate question-answer pairs. BERT is used to rank the top

2022	Big Data Analytics	question-answer pairs.
Kumari <i>et al.</i> , 2022	International Conference on Cloud Computing, Data Science & Engineering	T5 was used to detect answers from texts and then use these answers to generate questions.
Maheen <i>et al.</i> , 2022	PeerJ Computer Science	BERT was used solely for text embeddings.
Malhar <i>et al.</i> , 2022	International Conference on Intelligent Computing and Control Systems	The T5 model was used to tokenize sentences, generate question-answer pair output from paragraph input, and then use the question-answer output to generate distractors for multiple-choice questions. BERT was used for question filtering and evaluation, as well as for text summarization to identify the most important sentences.
Mathur & Suchithra, 2022	International Conference on Computational Intelligence and Sustainable Engineering Solutions	BERT was used for text embeddings to extract key phrases (serving as answers), which were then used to generate questions and distractors by another NLP model (not a large language model).
Matsumori <i>et al.</i> , 2023	IEEE Access	BERT variants are used to predict masked tokens to create cloze questions.
Maurya & Desarkar, 2020	International Conference on Information & Knowledge Management	BERT was used to embed the text of input triplets in the format of <article, question, correct answer>. The embedded tokens were then passed to a sequence-to-sequence model to generate distractors.
Muse <i>et al.</i> , 2022	arxiv	T5 was utilized to generate questions with either pre-training or no pre-training.
Nguyen <i>et al.</i> , 2022	European Conference on Technology Enhanced Learning	T5 was used to generate questions with texts processed by different methods. GPT-3 was then used to evaluate the soundness of the generated questions.
Nittala <i>et al.</i> , 2022	Information and Communication Technology for Competitive Strategies	BERT was used for tokenization and word embedding. SCIBERT served as the generator, using the tokens of contexts and answers as input and also creating embeddings for these inputs.
Offerijns <i>et al.</i> , 2020	arxiv	In the first phase, GPT-2 generated questions using context and answers as input, and in the second phase, it generated distractors using context, answers, and questions as input. BERT was used to filter questions for answerability and coherence (question filtering).
Pochiraju & al., 2023	International Conference on Intelligent Computing and Control Systems	XLNet and BERT were used as text summarizers, with the output passed to other general NLP tools for question generation.
Qiu <i>et al.</i> , 2021	International Conference on Asian Language Processing	BERT was utilized for text embedding and question filtering. The distractor generation was still based on traditional methods such as random selection, cosine similarity, and Lowenstein distance.
Raina & Gales, 2022	arxiv	GPT-3 was used in a zero-shot manner, while T5 was used with context texts (without answers or key phrases) as input.
Ratcheva <i>et al.</i> , 2022	International Database Engineered Applications Symposium	T5 models were fine-tuned to generate questions with texts as input and with texts and answers combined as input.
Rodriguez-Torrealba <i>et al.</i> , 2022	Expert Systems with Applications	First, T5-small was used to extract answers based on input texts. Next, the sentence-answer pairs were fed into a T5-base model to generate questions.

Shan <i>et al.</i> , 2019	International Conference on Technologies and Applications of Artificial Intelligence	BERT was first used for sentence embedding of articles, enabling their entry into a transformer model to extract sentences corresponding to a given question and the answer (key phrase identification). Next, the question generation task is conducted using a sequence-to-sequence model with the extracted sentences as input and questions as output.
Shan <i>et al.</i> , 2022	Journal of Information Science and Engineering	BERT was used for text embedding, followed by question generation by a transformer model using the embedded tokens as input.
Shridhar <i>et al.</i> , 2022	arxiv	T5 was used with questions and contexts as input.
Srivastava & Goodman, 2021	arxiv	GPT-2 was used to generate new questions based on student performance and desired difficulty for the next item.
Steuer <i>et al.</i> , 2020	European Conference on Technology Enhanced Learning	Key phrase extraction was conducted before training GPT-2 to generate questions based on discussing the determined key phrases (a type of template-based generation).
Tsai <i>et al.</i> , 2021	International Conference on Computers in Education Conference	BERT was used to extract keywords from input textbooks, which were then used to construct important complete sentences. Subsequently, GPT-2 iteratively predicted the next word for question generation, using these complete sentences as input.
von Davier, 2019	arxiv	GPT-2 was utilized to generate distractors with QA inputs like “Q: What was A?” and “A: A was Y”, and to create question passages based on prompts.
Vu & Van Nguyen, 2022	Asian Conference on Intelligent Information and Database Systems	LLMs were trained to generate texts by employing reinforcement learning for attention selection in their model architecture.
Wang <i>et al.</i> 2021	Annual Meeting of the Association for Computational Linguistics	BERT served as a baseline model in their study, with no further details of implementation provided.
Wang <i>et al.</i> 2022	International Conference on Artificial Intelligence in Education	GPT-3 was employed using various prompting strategies for comparison. The first strategy involved using context and answer as input, while the second also utilized context and answer as input.
Wang <i>et al.</i> 2023	Education and Information Technologies	SentenceBERT was used to filter texts from slides and speech recognition, using the results as input for training T5, which was then used to generate answers. These answers were combined with the input texts to generate questions.
Wu, 2022	Master’s thesis	Prompts with masked tokens: “Question for answer A: [MASK] for context C”.
Xie <i>et al.</i> , 2021	IEEE/ACM Transactions on Audio, Speech, and Language Processing	T5 was used to generate questions with the filtered key sentences as input.
Zhang, 2023	arxiv	T5 was trained using texts and answers as input.
Zhao <i>et al.</i> , 2022	Annual Meeting of the Association for Computational Linguistics	First, BERT was used to extract the question type information. Second, this information was used as a signal for the BART model to summarize texts, which then passed to another BART model to generate questions. BERT was used solely for text embeddings.