Computers and Education: Artificial Intelligence 6 (2024) 100252

Contents lists available at ScienceDirect

Computers and Education: Artificial Intelligence

journal homepage: www.sciencedirect.com/journal/computers-and-education-artificial-intelligence

Automatic question-answer pairs generation using pre-trained large language models in higher education

Jintao Ling, Muhammad Afzaal *

Department of Computer and Systems Sciences, Stockholm University, Sweden

ARTICLE INFO

Keywords:
Pre-trained language model
Question-answer pairs generation
Higher education
Automatic evaluation
Real-educational evaluation

ABSTRACT

The process of manually generating question and answer (QA) pairs for assessments is known to be a time-consuming and energy-intensive task for teachers, specifically in higher education. Several studies have proposed various methods utilising pre-trained large language models for the generation of QA pairs. However, it is worth noting that these methods have primarily been evaluated on datasets that are not specifically educational in nature. Furthermore, the evaluation metrics and strategies employed in these studies differ significantly from those typically used in educational contexts. The present discourse fails to present a compelling case regarding the efficacy and practicality of stated methods within the context of higher education. This study aimed to examine multiple QA pairs generation approaches in relation to their performance and the efficacy and constraints within the context of higher education. The various approaches encompassed in this study comprise pipeline, joint, multi-task approach. The performance of these approaches under consideration was assessed on three datasets related to distinct courses. The evaluation integrates three automated methods, teacher assessments, and real-world educational evaluations to provide a comprehensive analysis. The comparison of various approaches was conducted by directly assessing their performance using the average scores of different automatic metrics on three datasets. The results of the teachers and real educational evaluation indicate that the assessments generated were beneficial in enhancing the understanding of concepts and overall performance of students. The implications of the findings from this study hold significant importance in enhancing the efficacy of QA pair generation tools within the context of higher education.

1. Introduction

In the context of higher education, teachers face the significant challenge of creating assessments that are both effective and efficient in evaluating student knowledge Gholami & Morady Moghaddam, 2013; Jia et al., 2020). Traditional methods of manually crafting quizzes and exams are not only time-consuming but also detract from the valuable time that could be spent on teaching and engaging with students (Gholami & Morady Moghaddam, 2013; Klein & Nabi, 2019; Rodriguez-Torrealba et al., 2022b).

This situation highlights a pressing need for innovative solutions that can streamline the assessment creation process without sacrificing the quality of education.

This necessitates exploring the promising field of automated question-answer (QA) pair generation. Leveraging advancements in natural language processing and machine learning, particularly with pre-trained large language models (LLMs), offers a promising solution to streamline assessment development.(Alberti et al., 2019, pp. 6168–6173; Chan & Fan, 2019, pp. 154–162). Automated QA pair generation involves creating relevant questions and their corresponding answers from unstructured texts, aiming to streamline assessment development (Alberti et al., 2019, pp. 6168–6173; Chan & Fan, 2019, pp. 154–162). The significance of this research lies in its potential to revolutionize how assessments are developed, making the process more efficient and opening the door to more personalized and adaptive learning experiences. By evaluating different QA pair generation methodologies—namely the pipeline, joint, and multi-task learning approaches—this study aims to identify the most effective strategies for integrating this technology into higher education. This is not merely an academic exercise; it has the potential to significantly impact educational practices, enhancing the way teachers design assessments and ultimately improving student learning outcomes.

However, the aforementioned methodologies were assessed and validated using separate datasets and evaluation metrics. When

* Corresponding author.
E-mail address: muhammad.afzaal@dsv.su.se (M. Afzaal).

https://doi.org/10.1016/j.caeai.2024.100252
Received 29 November 2023; Received in revised form 24 April 2024; Accepted 8 June 2024
Available online 10 June 2024

considering the application of these methodologies within the field of education, it becomes challenging to offer a definitive recommendation for approach selection due to the difficulty in directly comparing their efficacy. While certain literature sources acknowledge the potential use of their approaches in an educational context, there is a lack of empirical evaluation of these approaches in real-world educational settings. Additionally, the datasets (e.g., SQuAD dataset is extensively utilised and has over 100,000 question-answer pairs derived from a collection of Wikipedia articles) that used to generate pairs specifically designed for reading comprehension (Rajpurkar et al., 2016). However, using datasets like SQuAD for educational efficacy assessment is inappropriate, as they were not specifically designed for educational purposes (Lelkes et al., 2021). As a response, diverse methodologies have been developed for QA pair generation, each holding potential for application in higher education.

In this context, we pose the following research questions.

- **Question 1:** How can we develop a robust framework to evaluate the effectiveness of different QA pairs generation approaches in the higher education context, ensuring that these technologies meet the specific needs of educational stakeholders?
- **Question 2:** What impact do automatically generated QA pairs have on students' academic performance and engagement in real educational environments?

The first research question underscores the necessity of a tailored evaluation that transcends mere technical performance, integrating pedagogical considerations to truly assess the value of these technologies in educational settings. While the second research question is critical as it seeks to directly link the use of innovative assessment technologies with measurable outcomes in student learning and engagement, thus providing a compelling argument for the integration of such tools in educational practice.

By proposing a methodology that evaluates the effectiveness of different QA pair generation approaches (pipeline, joint, and the multi-task approaches) using pretrained large language models (LLMs) across three benchmark datasets derived from real educational materials, this paper aims to fill the aforementioned gaps. These datasets reflect different courses that are taught at University and were generated specifically for evaluating different methodologies. The evaluation process involves automatic, teachers and real-educational settings within the context of higher education. Our findings reveal that the multi-task learning approach, particularly when utilising the T5 LLM, outperforms its counterparts, demonstrating a notable positive impact on student academic performance. Teachers expressed high satisfaction with the question accuracy and relevance, acknowledging the technical competence and educational utility of the QA pairs, while noting the need for improvements in understandability and consistency across different courses. Moreover, our investigation shows that the use of generated QA pairs positively influences students' academic performance. Specifically, it was observed that students engaging more frequently with QA pair-based assessments tended to perform better in their final examinations, indicating the significant benefits of integrating these technologies into educational assessments.

2. Background

The utilisation of pre-trained language models has yielded remarkable achievements in the domain of natural language processing (NLP). Consequently, numerous methodologies have been put forth to address the task of generating question-answer pairs, leveraging the capabilities of these pre-trained language models. These methodologies are not merely technical achievements; they have practical implications for educational practices, particularly in crafting personalized and adaptive learning experiences (Yi et al., 2021).

These approaches have demonstrated substantial enhancements in performance, as evidenced by empirical evaluations (Rodriguez-Torrealba et al., 2022b). The various approaches can be broadly classified into three distinct categories based on their architectural designs (Qu et al., 2021; Rodriguez-Torrealba et al., 2022b). These categories include the pipeline approach, joint learning approach, and multi-task learning approach.

2.1. Pipeline approach

The pipeline approach is a straightforward methodology in which the processes of question generation and answer generation are executed sequentially. In a recent study (Rodriguez-Torrealba et al., 2022b), the authors introduced a novel processing pipeline that leverages the T5 language model to generate question-answer pairs. This method's superior performance in generating coherent and contextually relevant QA pairs suggests its potential to enhance learning by providing more engaging and challenging materials in higher education settings (Johnson et al., 2024). The model, known as the QAP Model (Question/Answer Pairs Model), leverages the inherent duality between question generation (QG) and QA to achieve its objectives. The T5 model is utilised for generating both questions and answers, with a subsequent model employing distractors to provide incorrect answer options, which enhances the learning challenge.

In another study (Alberti et al., 2019, pp. 6168–6173), authors introduced a roundtrip consistency mechanism to further enhance the coherence of the generated QA pairs. This mechanism involves a detailed comparison between initial and subsequent responses to ensure coherence, excluding pairs that do not meet the criteria. In the context of education, it is crucial to tailor the generation of student questions to meet educational needs rather than solely assessing comprehension (Yao et al., 2022). The FAIRYTABLEQA dataset, annotated by domain experts, is utilised to refine the rules for answer extraction, although limitations exist in covering the broad spectrum of educational subjects. Researchers also propose integrating a text summarization module within the pipeline approach to streamline the extraction of crucial information and the formulation of questions and answers (Gabajiwala et al., 2022). This optimized approach utilizes advanced tools like ConceptNet, WordNet, and Sense2vec to ensure relevance and accuracy in the generated educational content.

The straightforward pipeline approach using LLMs like T5 in higher education allows for ease of implementation, which is beneficial for teachers without deep technical expertise. It facilitates the customization and scalability of educational content, making it adaptable to various class sizes and curriculum needs (Kurdi et al., 2020). Moreover, the modular nature of the pipeline permits iterative enhancements based on feedback, contributing to continuous improvement in teaching methodologies and student learning outcomes. This methodological simplicity thus not only democratizes access to advanced AI technologies in educational settings but also enhances the overall educational experience by providing reliable and adaptable tools for both teaching and assessment (H. C. Wang et al., 2023).

2.2. Joint learning approach

In the context of a joint model, the process of generating both the question and answer occurs iteratively. The interdependence between question generation and question answering presents a promising opportunity for enhancing performance and reducing the reliance on annotated data (Klein & Nabi, 2019). The pipeline approach exhibits suboptimal performance in the task of extracting the most suitable QA pairs from textual data. This is primarily due to the disregard for the interdependence between question generation and answer extraction, resulting in the potential generation of incompatible QA pairs (Cui et al., 2021).

A novel approach was introduced (Klein & Nabi, 2019) that integrates the Transformers decoder GPT-2 model with the Transformers

encoder BERT, with the aim of facilitating collaborative learning in the context of question-answering and question generation. The training of the model is conducted through the utilisation of an endto-end schema, wherein the dual tasks of question-answering and question generation are considered concurrently, rather than being approached as sequential tasks. The utilisation of a ground-truth answer and accompanying context text is employed as a means to generate a question through the utilisation of GPT-2. The subsequent step involves providing the pre-trained BERT model with the SQuAD text, which encompasses the question that was formulated, in order to facilitate the extraction of the answer span. In the event that BERT fails to accurately predict the correct answer as indicated by the annotations, the resulting loss of information will be propagated backwards to GPT-2.

In a similar vein, a novel unified framework for the generation of QA pairs was proposed (Qu et al., 2021). Based on the available evidence, it can be inferred that the process of sequential generation can be effectively simplified by transforming it into the task of generating joint question-answer pairs. Within the confines of this particular framework, the tasks of question generation and keyphrase extraction have been intricately intertwined, forming a dual task. The optimisation process involves leveraging the advantages of each task in a mutually reinforcing manner, with iterative refinement being a key component. The process of generating answers will be facilitated by utilising the extracted keyphrase as a guiding factor.

The joint model approach with BERT or GPT is transformative, integrating question and answer generation to reflect the interconnected nature of learning. This method enhances educational practices by providing immediate, context-aware feedback and generating precisely aligned question-answer pairs, leading to more accurate student assessments (Rodriguez-Torrealba et al., 2022a). It also allows for the efficient creation of customized educational content, significantly reducing the workload on educators and enabling them to focus on more personalized teaching. Moreover, its ability to automate content generation makes it ideal for scalable applications such as online learning platforms and MOOCs, supporting large-scale educational initiatives with consistent, high-quality content (Kurdi et al., 2020).

2.3. Multi-task learning approach

The multi-task model utilizes a shared encoder to process inputs for both question and answer generation, enabling mutual learning and synergy between these tasks. This integration blurs the lines between generating questions and answers, highlighting their interdependence and enhancing their effectiveness (Cui et al., 2021). One of the primary challenges encountered in the pipeline approach is the generation of incompatible QA pairs in an independent manner, without considering the underlying relationships between them. The OneStop model is being proposed as a solution to address these issues. It leverages a sequence-to-sequence transformer architecture, featuring a bidirectional encoder and an autoregressive decoder, to generate questionanswer pairs using a multi-task learning approach (Cui et al., 2021). Documents are inputted into the encoder, and then the decoder, employing cross-attention with the encoder's outputs, generates questions and predicts answer spans. This interplay between the document and the generated content is crucial for producing accurate and compatible responses, thereby achieving optimal results in the question and answer generation processes.

In contrast to the sharing of the encoding model on question generation and answer generation tasks, the multi-task-based T5 model demonstrates a more direct approach in modelling these tasks (Zhong et al., 2022). In the multi-task setting, the T5 model undergoes fine-tuning on three distinct tasks: question answering, question generation, and answer extraction. For the question answering task, the model takes in a context and question pair as input. In the question generation task, the model utilizes answerhighlighted context as input. Lastly, for the answer extraction task, the model employs sentence-highlighted context as input. The performance of the multi-task-based model is observed to be superior when evaluating its results in comparison to the T5 model that has been fine-tuned under a single-task setting, as reported in reference (Akyon et al., 2022).

In higher education, the use of QA pairs based assessments using LLMs like BERT and T5, employing approaches such as pipeline, joint learning, and multi-task learning, significantly enhances both teaching and learning experiences. For students, these models facilitate personalized learning through adaptive assessments and provide immediate, detailed feedback, which is crucial for effective learning (Zhong et al., 2022). They also help in creating fairer and more accurate evaluations by minimizing human biases and ensuring consistency across assessments.The direct impact on student learning is profound, as these assessments help in reinforcing knowledge through repeated and tailored practice, leading to improved understanding and retention (Mazidi & Nielsen, 2014). For teachers, LLMs offer considerable benefits by automating the generation of test materials, thus saving time and effort that can be redirected towards interactive teaching. Additionally, the data-driven insights generated by these models support informed pedagogical decisions and curriculum development, enhancing educational outcomes. The superior performance of LLMs such as T5 model in higher education settings means that educational practices can become more adaptive, personalized, and inclusive, enhancing both the quality and accessibility of education (Kurdi et al., 2020).

2.4. Research gap

In conclusion, it is worth noting that the existing methodologies for generating QA pairs using pre-trained language models can be classified into three distinct categories: the pipeline method, the joint model, and the multi-task learning model. The approaches that have been reviewed are organised and presented in a structured manner in table 0.8. The utilisation of multiple datasets for the generation of QA pairs is evident in the aforementioned methods. Furthermore, it is important to note that the metrics employed for the purpose of automatically evaluating text generation were found to be inconsistent, lacking a standardised set of metrics for comprehensive evaluation.

The current landscape of QA pairs generation methodologies has witnessed the emergence of various proposed approaches. However, a notable gap in the existing literature relates to the absence of comprehensive performance evaluations conducted on real-world education datasets. The performance of these approaches has been assessed through the utilisation of various datasets and evaluation metrics, thereby posing a challenge in terms of comparing their respective outcomes. Furthermore, it should be noted that the datasets provided in table 0.8 lack a clear source from real-world educational settings. Consequently, the extent to which these datasets can be applied to the field of education remains uncertain. Hence, it is imperative to conduct a comprehensive assessment of these methodologies using a universally accepted dataset and standardised evaluation techniques in order to ascertain their efficacy in the realm of education. Furthermore, it is imperative to incorporate real-world evaluations that involve educators with substantial teaching experience in the field of higher education. These evaluations are crucial in order to comprehensively assess the influence and efficacy of these approaches.

Moreover, there is a notable absence of empirical studies directly investigating the impact of automatically generated QA pairs on students' academic performance and engagement. This gap highlights the need for targeted research that evaluates how these technologies affect learning outcomes in actual classroom settings. Such studies are essential to understand whether and how the use of automated QA pairs can enhance educational practices, particularly in terms of improving student engagement and academic achievement.

3. Methodology

This section outlines a methodology for assessing the efficacy of various approaches to generating QA pairs and different LLMs within the context of higher education. Another objective is to evaluate the influence of these approaches and LLMs on students' academic performance. The methodology consists of three main phases. The initial phase of data collection entails gathering fin-tune datasets for the purpose of fine-tuning LLMs. Subsequently, benchmark datasets are generated to facilitate the evaluation process. In the second phase of the experiment, three distinct approaches for generating QA pairs will be selected. These approaches will be combined with LLMs that will be fine-tuned using the fine-tune datasets that have been collected. In the evaluation phase, the selected approaches and LLMs are assessed on benchmark datasets using various accuracy measures. The goal is to determine the most suitable group, consisting of both the approach and LLM, for real-educational evaluation. The selected cohort is subsequently employed to generate MCQ assessments, which are employed in the context of real-world educational evaluations. These evaluations seek to evaluate the influence of the assessments on students' academic achievements. The subsequent sections will provide a comprehensive analysis of the various stages involved in the proposed methodology.

3.1. Data collection

This section introduces and describes two distinct categories of datasets that were gathered for the purpose of this research work. Initially, it is important to highlight the significance of fine-tuning datasets in the process of refining pre-trained language models. These datasets are specifically designed to improve the performance of these models in tasks that are specific to particular domains. Secondly, benchmark datasets, derived from educational materials through manual curation, are predominantly utilised for the evaluation of performance in fine-tuned models.

3.1.1. Fine-tuning datasets

The approach for QA pairs and DG generation were fine-tuned using the SQuAD and DG-RACE datasets, respectively. The utilisation of multiple datasets was observed in the process of fine-tuning the chosen methods for generating QA pairs, as indicated in Table 1. The selected datasets for this study exhibit distinct characteristics and are applicable in various scenarios.

The Stanford Question Answering Dataset (SQuAD) uniquely comprises passages from Wikipedia articles and corresponding question-answer pairs, highlighting its utility for evaluating reading comprehension in LLMs models. Each entry features a passage and a set of questions formulated from the text, where answers are exact text spans within the passage (as shown in Fig. 1). This structure aims to challenge models across a spectrum of topics, enforcing a rigorous test of their natural language processing capabilities. Documented by Rajpurkar et al. (Rajpurkar et al., 2016), the dataset emphasizes creating one high-quality question-answer pair per passage, ensuring precise model evaluation. SQuAD's design, leveraging a wide subject range, from history to science, serves as a comprehensive benchmark for advancing question-answering systems and enhancing machine learning techniques.

The DG-RACE dataset, pivotal for enhancing distractor generation (DG) techniques, serves as the foundation for fine-tuning pre-trained language models aimed at creating distractors for multiple-choice questions (MCQs) in reading comprehension exams (Gao et al., 2018; Rodriguez-Torrealba et al., 2022b). It is designed not just to offer incorrect alternatives but to ensure these distractors are convincingly plausible, thereby enriching the realism and educational value of practice exams (As showin in Fig. 2). The dataset leverages a wide array of

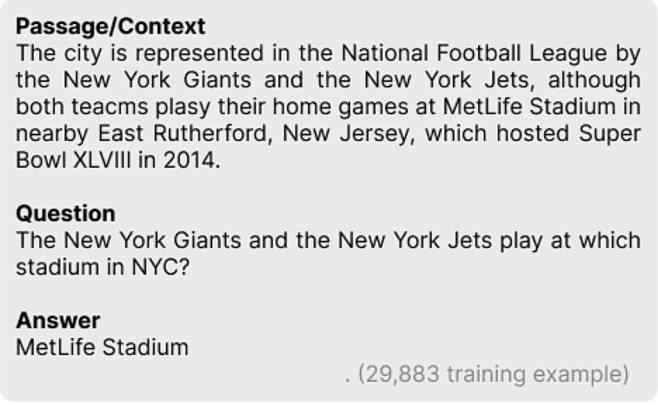

Fig. 1. SQuAD dataset example.

Table 1
The summarization of the proposed method.

Model	Dataset	Question generation	Purpose: Answer generation	Distractor generation	Pre-trained Language model	Method type	Automatic Evaluation	Expert Evaluation	Evaluation in a realworld setting
QAPModel (Rodriguez-Torrealba et al., 2022)	DG-RACE	✔	✔	✔	T5	Pipeline	cosine similarity	×	×
QACGModel (Alberti et al., 2019, pp. 6168–6173)	SQuAD, natural Questions (NQ)	✔	✔	×	BERT	Pipeline	EM, F1	✔	×
QAG system (Yao et al., 2022)	FairytaleQA	✔	✔	×	BART	Pipeline	ROUGE	✔	×
Question Generation and Answering (Klein & Nabi, 2019)	SQuAD	✔	×	×	GPT-2, BERT	Joint Learning	BLUE, ROGUE	×	×
QAG framework (Qu et al., 2021)	SQuAD, RACE	✔	✔	×	ProphetNet	Joint Learning	BLEU, ROUGE, METEOR	✔	×
OneStop Model (Cui et al., 2021)	SQuAD, NewsQA, DuReader	✔	✔	×	BART	Multitask Learning	BLEU, ROUGE	✔	×
Multi-task mT5 model (Akyon et al., 2022)	TQuADv2	✔	✔	×	mT5	Multitask learning	BLEU, METEOR, ROUGE	×	×

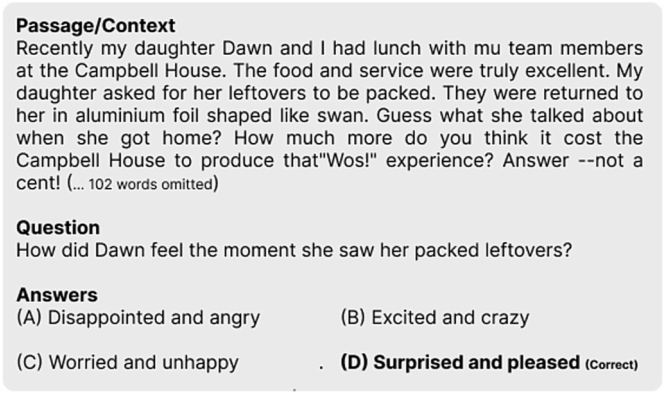

Fig. 2. DG-RACE dataset example.

texts, drawing from subjects commonly found in academic assessments to ensure a broad applicability. By focusing on the generation of high-quality, contextually relevant distractors, DG-RACE aims to simulate a genuine exam experience, testing students' comprehension skills effectively and preparing LLMs based systems to support nuanced educational needs.

3.1.2. Benchmark datasets

To ensure uniformity and high educational standards, we recognized the need for a benchmark dataset tailored explicitly for educational purposes, distinguishing it from datasets like RACE and SQuAD, which are not primarily designed with educational objectives in mind. This dedicated dataset adheres to strict criteria, emphasizing the creation of questions that include a passage/context, one correct answer, and three carefully crafted distractors to simulate real-world examination conditions effectively (as shown in Fig. 3). Inspired by the methodologies proposed by Torres (Torres et al., 2009), our approach to building this benchmark datasets involved detailed guidelines, as documented in Appendix A.

We selected three courses for this endeavor: a bachelor-level course in programming 1 and two master-level courses in Big Data with NoSQL and Enterprise Computing and ERP Systems. These courses were chosen to develop benchmark datasets, ensuring a wide range of coverage that reflects the critical areas of contemporary computer science education. The annotation process was meticulously carried out with the collaboration of six annotators, comprising teaching assistants, to ensure a robust evaluation across three distinct courses. For each course, we carefully selected two teaching assistants for each course, prioritizing their expertise and familiarity with the course content. These assistants were tasked with selecting and refining material directly from academic textbooks, with a specific focus on ensuring the relevance and challenge of each question and its set of distractors. To further enhance the quality and applicability of the dataset, we adopted a consensus-based approach. Only those questions where both the course instructor and the teaching assistant reached agreement were included. This rigorous selection criterion ensured that every question of the dataset not only aligns closely with academic standards but also effectively tests students' knowledge and critical thinking skills in a manner that is both challenging and educationally valuable.

3.2. Experiment

Based on our analysis of the existing literature, it can be inferred that the approaches employed for QA pairs can be categorised into three distinct types: the pipeline, joint, and the multi-task learning approach. However, there are number of approaches that fall in theses types. The experimental design involved the selection of approaches for this study which is based on four conditions. First, the approach must satisfy the specific requirements of the educational context, which includes the generation of questions, corresponding answers. Second, the approach has the ability to generate the distractors will be considered preferentially. Third, the approach has been mentioned the applicability in the education context. Lastly, the approach must be designed specifically for the English language. Based on thses conditions three approaches were selected (Cui et al., 2021; Qu et al., 2021; Rodriguez-Torrealba et al., 2022b).

However, the selection of the pre-trained language models is a crucial factor that can significantly impact the performance of these approaches. It means that the different pre-trained language models have been shown to exhibit bias based on the training data used and model architectures. To demonstrate the performance of the proposed approaches explicitly, different pre-trained language models were selected and accessed respectively to exploit the patience of each approach and minimize the potential bias of different pre-trained language models. As a result, a combination of pre-trained language models with comparable architectures to those employed in the original research were used in conjunction with the selected approaches to generate QA pairs.

In this study, we meticulously selected pre-trained language models such as T5, ProphetNet, and BART based on specific criteria to ensure alignment with our research goals on QA pairs generation (Liu et al.,

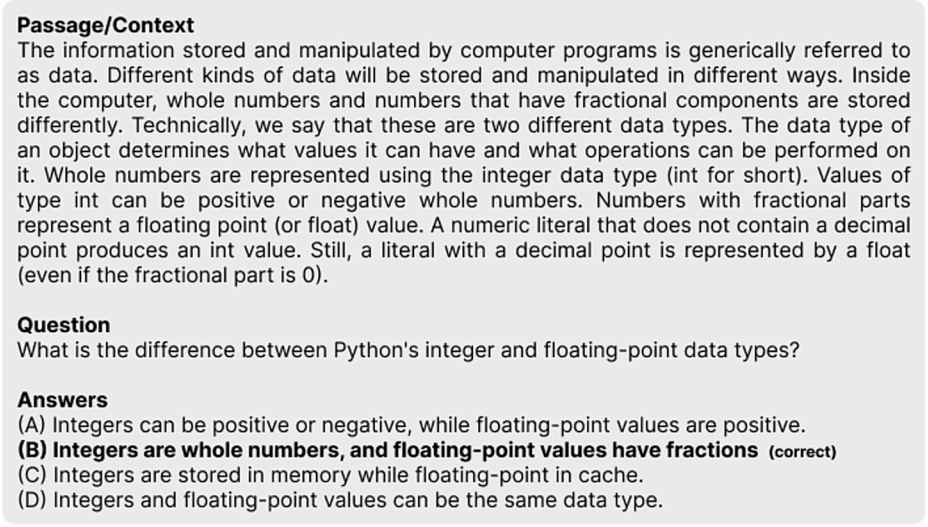

Fig. 3. Benchmark dataset example.

2019; Qi et al., 2020; Raffel et al., 2020; H. Wang et al., 2022). Key considerations included each model's architectural innovation, like T5's text-to-text framework, ProphetNet's n-gram prediction, and BART's hybrid approach, ensuring advanced comprehension and generation capabilities. Performance on benchmark datasets was crucial, guiding us to models with demonstrated excellence in NLP tasks. Flexibility for fine-tuning allowed us to adapt models to our unique dataset, enhancing their effectiveness. The selection was further influenced by the level of community support and the availability of resources, ensuring ease of implementation and potential for collaboration. Lastly, engagement with models at the forefront of NLP research promised that our work would remain cutting-edge. This thoughtful selection process guaranteed that the models we chose were not just tools but integral to pushing the boundaries of QA pairs generation research.

Thus, these three pre-trained language models can serve as substitutes for each other in the selected approaches. In summary, Table .2 shows the combinations of the QA pairs generation approaches and pre-trained LMs. The experiments were carried out using the SQuAD dataset.

Additionally, the distractors were an essential part of the MCQ, except the correct question and answer. A T5-based distractor generation (DG) model with the question-answer pairs and the context paragraph was implemented to generate distractors for the generated QA pairs from all selected approaches to form a complete MCQ (Rodriguez-Torrealba et al., 2022b), which refers to the part of the pipeline approach.

3.2.1. Fine-tuning

The fine-tuning stage started with the data preprocessing. Then the processed data was passed to the QA generation models that have been shown in Table .2. Totally nine fine-tuned models were obtained.The data preprocessing method varies across different QA pairs generation approaches. For each approach, we will adhere to the original data processing strategy used in the source material. After the preparation of the data, the text will be given to the tokenizer of each pre-trained language model. Tokenization means breaking the natural language into chunks of tokens considered discrete elements (Liu et al., 2019). The subsequent step involves representing the text as a vector based on the frequency of tokens in the documents, thereby converting the text into a numerical data structure that can be used as input for the models.

In the process of fine-tuning our model, we conducted all our experiments using the powerful NVIDIA V100 GPU on the Google Cloud platform. We started with an initial learning rate of 10e5 and implemented a warm-up strategy to make sure the learning process stays stable. The batch size was aligned with 8. Our main task was predicting the next token, which is a classic challenge in building models that generate text. To optimize the model's parameters, we used the cross-entropy loss function along with the AdaW optimizer.

3.3. Evaluation

During the evaluation stage, the benchmark dataset is utilised as input for the fine-tuned models in order to generate QA pairs and distractors. Subsequently, the performance of each approach will be assessed through a comprehensive evaluation process involving automated, teachers and real-world evaluations. The evaluation employed to assess the efficacy of the generated QA pairs derived from diverse approaches. The objective of this evaluation is to assess the efficacy of the generated QA pairs in the context of educational applications.

Table 2
The Groups of the QA pairs generation methods and pre-trained LMs.

	Pipeline approach	Approach + LLM	Multi-task learning approach
		Joint learning approach	
T5	Pipeline + T5	Joint + T5	Multi-task + T5
BART	Pipeline + BART	Joint + BART	Multi-task + BART
ProphetNet	Pipeline + ProphetNet	Joint + BART	Multi-task ProphetNet

3.3.1. Automatic evaluation

In the context of automatic evaluation, it was necessary to establish nine distinct groups (as presented in Table 2) for the purpose of generating QA pairs since we are using three approaches along with three large language models. For example, in the context of the pipeline approach, three distinct groups were formulated, namely pipeline + T5, pipeline + Bart, and pipeline + ProphetNet.

The evaluation of these combinations involved the utilisation of established evaluation metrics, namely BLEU, METEOR, and ROUGE, to quantitatively assess their performance and derive corresponding scores. The computation of the average score is subsequently performed for every QA pair generated by each group. Based on the findings obtained from the evaluation results, it was deemed necessary to select the three groups (one group for each course) that exhibited the highest level of performance within all approaches for the purpose of conducting real-world educational evaluation. This decision was made in light of the fact that evaluating all nine groups would entail a significant investment of time and effort. While automatic evaluation metrics have proven to be valuable in objectively assessing the quality of generated text and reference text, there remains a gap in their ability to evaluate the syntactic and semantic aspects of generated QA pairs (Qu et al., 2021).

3.3.2. Teachers Evaluation

To conduct a thorough assessment of the quality and educational utility of the generated QA pairs, it is crucial to include feedback from teachers. This approach aligns with the methodology proposed by (Qu et al., 2021), emphasizing the importance of teacher insights in evaluating QA pairs' effectiveness in real-world educational contexts. To this end, three teachers, each representing a distinct course of study, were carefully selected to participate in the creation of assessments and to provide detailed feedback through in-depth interviews. These interviews were designed to critically assess the quality and utility of these QA pairs.

After that teachers were tasked with organizing the QA pairs and associated distractors into a series of Multiple Choice Questions (MCQs). This step was undertaken to facilitate the creation of assessments or exercises tailored to the educational content. A total of 48 assessments were crafted, with a detailed allocation of 19 assessments for the programming 1 course, 16 for the Big Data course, and 13 for the Enterprise Computing and ERP Systems course, reflecting the specific content and learning objectives of each course. Subsequent to the assessment creation, in-person semi-structured interviews were conducted with each teacher to gauge the assessments' quality. These interviews were meticulously recorded, and significant insights were documented by the interviewers. The recordings and key observations were subsequently compiled and subjected to a thematic analysis, aiming to distill and interpret the essence of the interview discussions, thereby providing a nuanced understanding of the educational value and potential areas for improvement of the QA pairs.

3.3.3. Real-world educational evaluation

In the stage of real-world educational evaluation, the objective was to assess the influence of the generated assessments on the academic performance of students across a range of courses. Therefore, half of the students in each course were randomly selected to deliver the assessments, and the remainder were not allowed to interact with the assessments. At the finish of the courses, a comprehensive data collection and analysis process was implemented in order to assess the influence of the delivered assessments on the academic performance of the students.

In terms of data collection, data related to the academic performance of students, as well as their efforts in completing the assessments, were gathered for the purpose of this study. This encompassed the final

examination scores of each student, as well as the quantity of assessment attempts made by students throughout the duration of the course. In terms of data analysis, a pair of statistical tests were conducted in order to assess the influence of the assessments on the academic performance of the students. In order to determine the difference in academic performance between students who engaged in the assessments and those who did not, a statistical analysis employing an independent samples *t*-test was conducted. In order to investigate the potential association between assessment attempts and students' academic achievement, this research employed Pearson and Spearman's correlation analyses.

4. Results

In this section, we present the results of the experiments conducted to evaluate the effectiveness of approaches. These approaches were evaluated in terms of its ability to correctly generate QA pairs and and how they affected students' academic success.

4.1. Results of automatic evaluation

The results of the automatic evaluation are presented in Table .3, Table .4 and Table .5 for the benchmark datasets of each group (Approach + model).

These tables show the results of various automatic evaluation metrics, including BLEU-1, BLEU-2, BLEU-3, BLEU-4, ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-Lsum, and METEOR, for three different pipeline groups (Pipeline + T5, Pipeline + BART, Pipeline + ProphetNet), three different joint groups (Joint + T5, Joint + BART, Joint + ProphetNet), and three different multi-task learning groups (Multi-task + T5, Multi-task + BART, Multi-task + ProphetNet). The values in tables represent the average score of each metric across all the QA pairs in the benchmark datasets. The higher the score, the better the performance of the group in generating high-quality answers. For example, the BLEU-1 score for the Pipeline + T5 group is 36.15, which means that on average, this group generates answers that match 36.15% of the unigrams in the reference answers.

In terms of approaches, it can be observed that the multi-task approach outperforms both the pipeline and joint approaches across all benchmark datasets representing three distinct courses. Notably, in the context of Programming 1, the pipeline method exhibits performance levels that are comparable to those of the multi-task method. Conversely, the joint approach lags behind both the multi-task and pipeline approaches. In the case of the enterprise modelling and big data course, the multi-task approach displayed the highest performance, while the joint and pipeline methods demonstrated comparable performance levels, with the joint model marginally outperforming the pipeline method.

In the context of LLMs, it has been noted that T5 demonstrates enhanced performance in comparison to other LLMs across a range of methodologies, including pipeline, joint, and multi-task approaches. Furthermore, the findings of this study indicate that across all three courses, T5 consistently demonstrated its efficacy. However, it is worth noting that within the specific context of the Programming 1 course, BART exhibited superior performance when utilising a multitask approach compared to other LLMs. Nevertheless, it is important to highlight that BART's performance, even in this scenario, was still inferior to that of T5 with a pipeline approach.

Based on the findings presented in the tabulated data (Tables 3 and 4 and Table .5) and following the established procedures for real educational evaluation (section 3.3.2), it is considered appropriate to exclusively utilise the most favourable outcome obtained from each respective group for the purpose of generating QA pairs. Henceforth, it has been determined that the QA pairs generated by the Pipeline + T5 group will be selected for the Programming 1 course. Similarly, for the big data and enterprise modelling courses, QA pairs obtained from the Multi-task + T5 group will be chosen.

4.2. Results of Teachers Evaluation

The thematic analysis of interview data on the evaluation of QA pairs identified five themes: correctness, understandability, difficulty level, knowledge impact, and utility impact (as shown in Fig. 4).

4.2.1. Themes Definitions

Correctness assesses the factual accuracy of questions, answers, and distractors, ensuring they're relevant and appropriate. Understandability evaluates the clarity and accessibility of language and presentation, making sure the content is easy to grasp. The difficulty level theme examines the complexity of the QA pairs, ensuring they are suitably challenging to effectively test student knowledge. Knowledge impact looks at the educational value of the QA pairs, from basic to advanced knowledge, assessing their scope in enhancing learning. Lastly, utility impact considers the overall usefulness of the QA pairs, evaluating how well they support learning outcomes, teaching effectiveness, and course quality. These themes together provide a comprehensive overview of the quality and educational significance of the QA pairs.

4.2.2. Results

The evaluation of the generated QA pairs across three course was conducted using a precise numerical scale, enabling a nuanced assessment of each dimension from low to high performance (as shown in Fig. 5). The evaluation results revealed high levels of teacher satisfaction with the questions' correctness, underscoring the QA pairs' accuracy, relevance, and appropriateness across different subjects. However, while the answers in the Programming 1 and Big Data courses were highly praised for correctness, the Enterprise Modelling course's results

Table 3
Results of automatic evaluation metrics for programming 1.

Metric	Models								
	Pipeline + T5	Pipeline + BART	Pipeline + ProphetNet	Joint + T5	Joint + BART	Joint + ProphetNet	Multitask + T5	Multitask + BART	Multitask + ProphetNet
BLEU-1	**36.14964029**	33.64153452	23.39791356	**30.33136522**	27.36548502	26.60121754	29.74910394	**30.98478783**	22.63743051
BLEU-2	**22.35608167**	20.90095886	11.69539345	**20.07960227**	16.95921112	16.65103447	19.2227172	**21.30069651**	12.57124711
BLEU-3	**15.79375148**	15.02378866	7.138231206	**15.2961294**	12.49040462	12.36952088	13.89763955	**16.39482792**	8.097514051
BLEU-4	**12.02456886**	11.82414574	5.032866744	**12.38900336**	10.12981764	10.06914535	10.60066968	**13.21698548**	5.763605075
ROUGE-1	**32.02495387**	30.63471727	21.11995874	**20.23716852**	16.99843471	15.73133455	**31.27625675**	31.18783253	23.80227394
ROUGE-2	**12.99585724**	12.75571406	5.401256265	**9.135552858**	6.925865541	6.635624063	14.38539526	**15.79856505**	6.857651749
ROUGE-L	**26.35652675**	23.9642481	15.96005056	**18.87403179**	15.88092056	14.33875987	25.05205253	**26.87081364**	18.81232257
ROUGE-Lsum	**26.36512806**	24.0792488	15.97624424	**18.87955415**	15.91985963	14.35597131	25.02281517	**26.85631877**	18.74756191
METEOR	**31.70870176**	29.47697555	25.37397826	**20.02663572**	17.75808498	19.41754084	**35.48587518**	34.20076926	32.69324558
Average Score	**23.97502333**	22.47792573	14.566210336	**18.36100481**	15.603120425	15.130016541	22.74361392	**24.09017744**	16.664761388

Table 4
Results of automatic evaluation metrics for big data.

Metric	Models								
	Pipeline + T5	Pipeline + BART	Pipeline + ProphetNet	Joint + T5	Joint + BART	Joint + ProphetNet	Multitask + T5	Multitask + BART	Multitask + ProphetNet
BLEU-1	**23.28076522**	18.2197417	20.79261672	**19.90801444**	18.93799357	16.43486243	**35.13891896**	32.38045738	22.77432712
BLEU-2	**16.50661274**	10.56219257	11.8668281	**13.27365872**	10.82287578	8.685232733	**26.96877127**	24.94276192	15.09960323
BLEU-3	**13.02351867**	6.741328941	8.339360611	**9.400936677**	7.010160883	5.780093828	**22.46610602**	21.15422776	11.43974895
BLEU-4	**10.90174311**	4.455817899	6.36770316	**6.913700666**	4.897609407	4.25681217	**19.43932548**	18.85497636	9.322526971
ROUGE-1	**27.28057174**	21.05013342	16.72537488	**24.29114063**	19.14611678	14.60227056	**29.45955894**	28.6514385	17.92825687
ROUGE-2	**13.83632432**	8.526407772	5.92788386	**12.08894383**	7.222503895	4.713001085	16.89802293	**17.49711636**	7.781688575
ROUGE-L	**26.93642493**	20.30977193	15.19678172	**23.79932539**	18.39886388	13.92933252	**28.2352558**	27.96082317	17.18851397
ROUGE-Lsum	**26.89281126**	20.34191829	15.16479933	**23.73454746**	18.42582878	13.97819757	**28.30308393**	27.80589989	17.25035657
METEOR	**29.23718962**	23.03642065	22.9554569	**27.68654117**	22.48453957	23.30335489	35.1346726	**35.42909757**	32.18946831
Average Score	**20.87732907**	14.804859241	13.704089476	**17.899645443**	14.149610283	11.742573087	**26.89374621**	26.07519988	16.774943396

Table 5
Results of automatic evaluation metrics for enterprise modelling.

Metric	Models								
	Pipeline + T5	Pipeline + BART	Pipeline + ProphetNet	Joint + T5	Joint + BART	Joint + ProphetNet	Multitask + T5	Multitask + BART	Multitask + ProphetNet
BLEU-1	24.89688773	**25.03339031**	20.66971081	**24.48834059**	23.1617379	21.11226446	**25.8684405**	25.3164557	16.21037464
BLEU-2	**16.27869483**	15.89980441	7.966343666	**16.46266515**	14.18031072	12.3615537	17.63112649	**17.63351073**	9.614484625
BLEU-3	**11.61481621**	11.38762279	3.953459766	**12.02050845**	9.649366618	8.273964464	12.93639526	**13.01625815**	6.195580785
BLEU-4	8.468833478	**8.593874495**	2.167852295	**8.815473961**	6.968471916	6.089207626	9.58311597	**9.638056692**	4.454534055
ROUGE-1	**23.41647287**	21.27903254	16.05618224	**24.7873407**	19.01818552	17.4928996	**27.8988838**	26.64783878	19.22624383
ROUGE-2	**10.30302322**	8.612521911	2.615157343	**11.81028214**	7.597897023	6.520264528	**14.84577498**	14.19347913	7.091385424
ROUGE-L	**22.07447606**	19.82906176	13.81788872	**23.70111652**	18.04581239	16.49337089	**26.27754562**	24.9953504	17.44916216
ROUGE-Lsum	**22.03421358**	19.73847621	13.8621216	**23.68067221**	18.07887164	16.50290914	**26.2163233**	25.05321598	17.46847245
METEOR	**22.65856622**	21.83261886	18.16281993	**23.98163752**	19.90253358	20.84400664	**30.26681373**	29.59724752	25.10922375
Average Score	**17.971776022**	16.911822587	11.030170708	**18.860893027**	15.178131923	13.965604561	**21.28049107**	20.676823676	13.646606858

indicated a medium level of satisfaction.

In the terms of understandability, the clarity and accessibility of the QA pairs were largely accepted, particularly the questions themselves, which were universally deemed highly understandable by all teachers. Nevertheless, the understandability of answers and distractors did not achieve unanimous high ratings, especially from the Enterprise Modelling and Programming 1 teachers, suggesting a need for improvement in how answers and distractors are articulated to ensure they are easily comprehensible.

The feedback on the difficulty level presented a mixed picture, indicating significant room for enhancement to better meet educational goals. While the Programming 1 course's teacher found the questions to be of a high difficulty level, aligning with desired standards, the responses regarding the difficulty levels of answers and distractors varied widely. This diversity in feedback highlights an imperative to recalibrate the challenge presented by the QA pairs, ensuring they are sufficiently demanding to stimulate student engagement and learning.

Teachers praised the QA pairs for their positive impact on imparting basic and concept-wise knowledge, with high evaluations across these areas. Yet, the advanced knowledge impact received lower ratings from the Programming 1 and Enterprise Modelling teachers, suggesting the necessity to delve deeper into complex topics and elevate the academic rigor of the QA pairs to foster a more comprehensive understanding among students.

Regarding utility impact, the QA pairs were recognized for their beneficial support towards students, teachers, and the overall course objectives, particularly in facilitating student learning and aiding teachers in assessing student comprehension. However, feedback indicated that improvements could be made in enhancing course support, as suggested by the medium level of satisfaction from the Enterprise Modelling teacher. This aspect underscores the potential to further refine the QA pairs to better align with and support the overarching goals of the courses.

In conclusion, while the method for generating quiz-style MCQs was deemed effective for educational purposes, especially in terms of correctness and understandability, the detailed feedback from teachers suggests several avenues for improvement. Enhancing the difficulty level, broadening the coverage of advanced knowledge, and bolstering course support emerged as key areas for future development.

4.3. Results of real-educational evaluation

In real-educational evaluation, created QA pairs were organized into series of MCQs in order to create assessments and delivered to students during the courses (section 3.3.2). In the end of the courses statistical analyses were carried out to investigate the effects of the assessments on student academic performing performance.

In this evaluation, the participants were divided into two distinct groups: Group 1, comprising individuals who were granted the opportunity to engage in assessments throughout the duration of the course, and Group 2, consisting of individuals who were not provided access to the assessments. Following the completion of the aforementioned tasks, a series of statistical analyses were conducted. The initial analysis aimed to determine any differences in academic performance between students who engaged in the assessments and those who did not. To achieve this, an independent samples *t*-test was employed. In order to investigate the

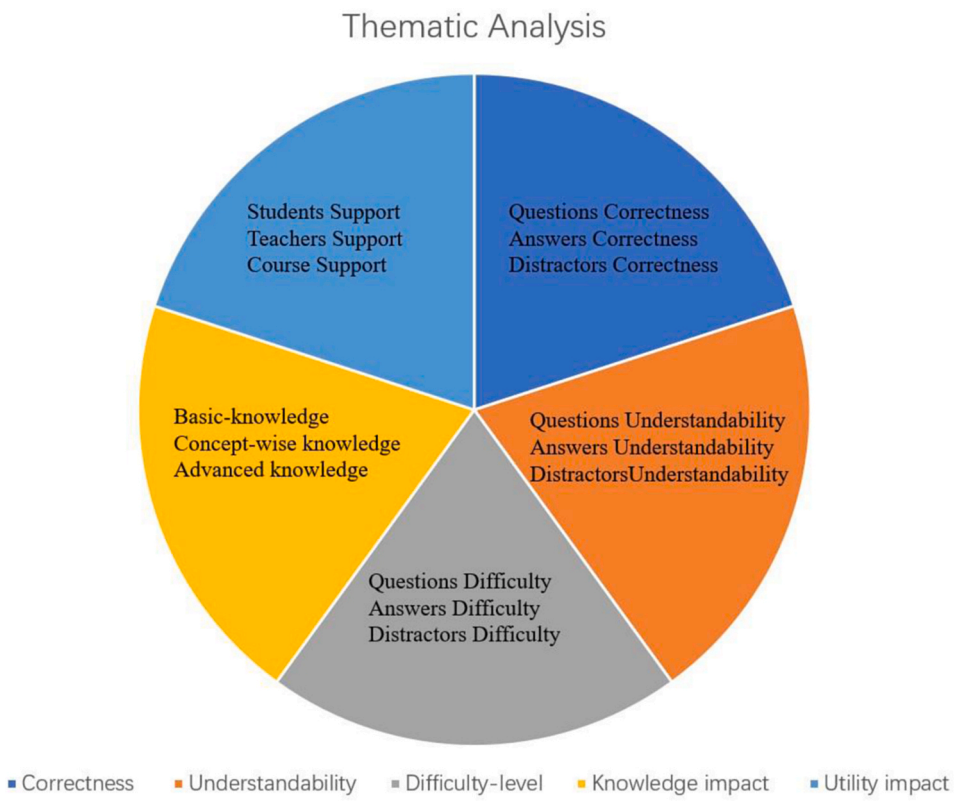

Fig. 4. Thematic analysis.

potential association between assessment attempts and students' academic achievement, this research employed Pearson and Spearman's correlation analyses.

The results obtained from the *t*-test for the t and p-values (Table 7) showed a significant statistical difference between the groups in all three courses. The observed difference between groups is more pronounced in the context of the big data course when compared to the other two courses. Table 6 showed statistics of both groups in all three courses. As presented in Table 6 Group 1's mean in all three courses was greater than Group 2's. The greater mean for Group 1 demonstrates that participants who performed assessments during the courses performed better in the final exam than nonparticipants (group 2). The maximum scores for the final exams in the Programming 1, Big Data, and Enterprise Modelling courses are 300, 50, and 50, respectively.

The findings of the correlation analysis, as displayed in Table 8, revealed a favourable correlation between assessment attempts and student acadamic performance. According to the findings shown in Table 8, the highest correlation coefficient observed for both the Pearson and Spearman techniques was 0.67, specifically in regard to the Programming 1 course. The Enterprise Modelling course exhibited the lowest correlation of 0.59 when analysed using both techniques. However, in the Big Data course, the correlation shown a modest improvement compared to the Enterprise course, but it declined in comparison to the Programming 1 course. Based on the findings of this analysis, it can be inferred that there exists a positive correlation between the number of assessments attempted by students throughout the course and their performance on the final exam.

5. Discussion

This study implements three distinct approaches and LLMs for generating QA pairs, which were fine-tuned using the SquAD and RICE datasets. To assess their performance in educational settings, we developed benchmark datasets for three separate courses. These approaches were then evaluated on these benchmarks through a combination of automatic methods, teacher assessments, and real-world educational evaluations.

5.1. Findings

We began by addressing the first research question of how to develop a robust framework to evaluate the effectiveness of different QA pairs generation approaches in higher education, ensuring that these technologies meet the specific needs of educational stakeholders. Our results from automatic evaluations and teacher assessments provide a comprehensive view of the performance of these technologies across various educational contexts.

Our findings highlight that the multi-task approach generally outperforms the pipeline and joint methods in generating QA pairs, as evidenced by the high scores across multiple automatic evaluation metrics, particularly for the Programming and big data courses. This indicates a robust capability of the multi-task approach to produce quality QA pairs that align closely with reference answers in these courses. Additionally, the T5 model showed superior performance across all methods and was consistently effective, further supporting its use in our recommended framework for generating QA pairs.

However, when considering the application of these approaches in a real educational setting, the feedback from teachers provides essential insights that should influence the evaluation framework. Teachers reported high satisfaction with the accuracy and relevance of the questions generated, particularly praising the QA pairs for their correctness in the Programming 1 and Big Data courses. This suggests that the QA pairs generated are not only technically competent but also applicable and useful in an educational context. Despite these positives, the feedback also pointed out areas needing improvement. The medium satisfaction levels in the Enterprise Modelling course and the mixed feedback on the understandability of answers and distractors across courses highlight the need for adjustments in how these QA pairs are formulated. This

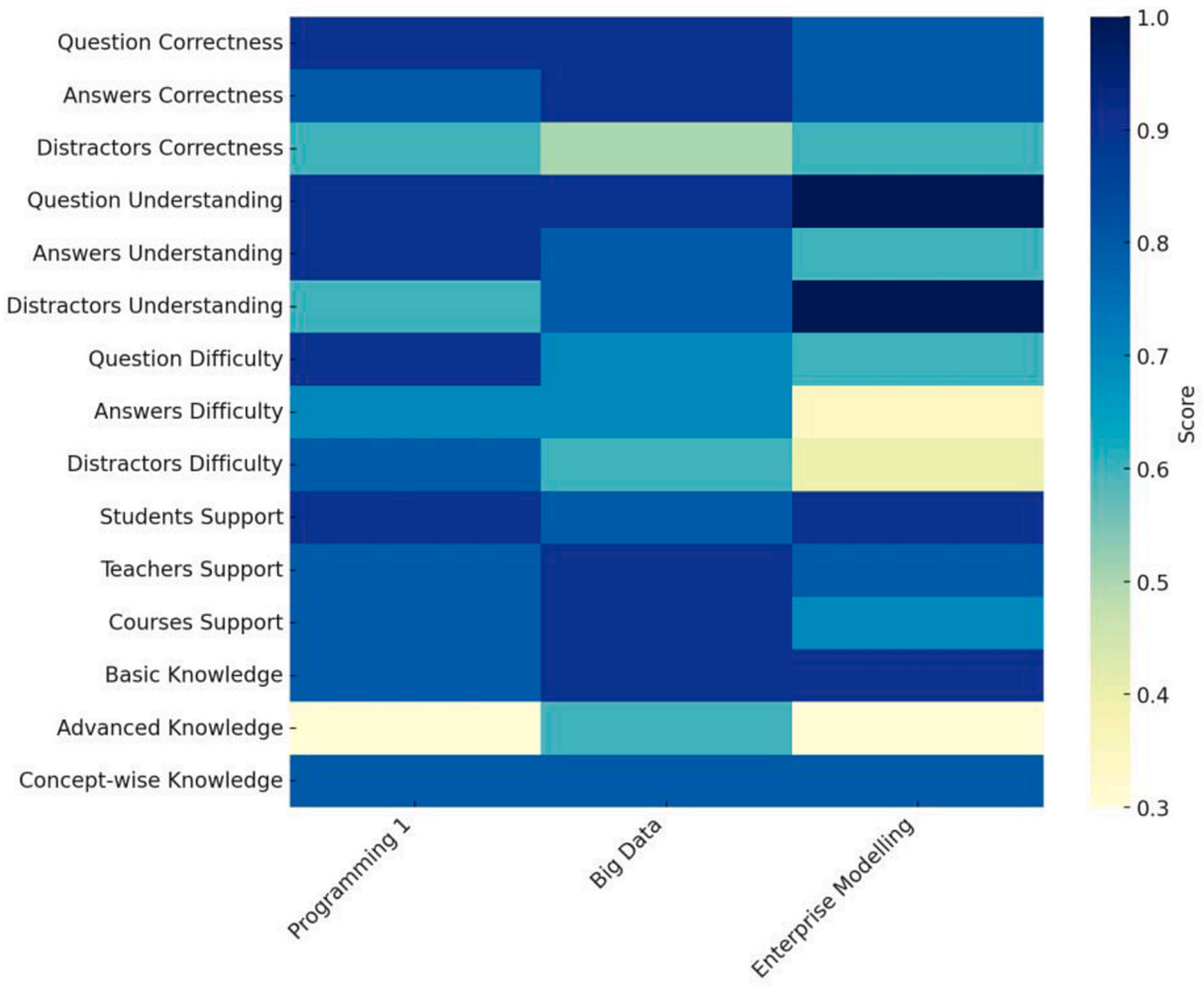

Fig. 5. Heatmap of thematic analysis of teachers evaluation.

Table 6
Statistics on groups of students.

Course	Group	N (users/ students)	Mean	Std. deviation
Programming 1	Group 1 (Assessments users)	120	231.51	59.68
	Group 2 (Non-Assessments users)	120	205.35	50.06
Big Data	Group 1 (Assessments users)	70	44.66	20.08
	Group 2 (Non-Assessments users)	70	32.37	8.27
Enterprise Modelling	Group 1 (Assessments users)	40	40.33	16.05
	Group 2 (Non-Assessments users)	40	33.53	9.34

Table 7
Results of the independent samples *t*-test.

Course	Sig.	t	df	Sig.(Two-Sided p)
Progamming 1	< 0.001	3.67	238	< 0.001
Big Data	< 0.001	4.73	138	< 0.001
Enterprise Modelling	< 0.001	2.31	78	0.023

variance underscores the importance of context and subject specificity in the development and application of QA generation tools.

Additionally, the teacher evaluations revealed that while the questions were effective at covering basic and conceptual knowledge, there was a noted need for deeper coverage of advanced topics, particularly in the Programming 1 and Enterprise Modelling courses. This feedback is crucial for refining the QA pairs to ensure they not only test but also enhance student understanding of more complex subject matter.

These findings lead us to the second research question, in which we aimed to examine the effects of the generated QA pairs on students' academic performance. To measure the effects of QA pairs we delivered MCQs assessments to students during the courses and at the end we performed two statistical analyses on final exam scores and students' assessment attempts. The results in Section 4.2 raised two important findings. First, the generated QA pairs positively impacted students' academic performance because students who were allowed to attempt assessments during the course achieved higher final exam scores than those who were not. Second, a positive correlation was found between the number of assessment attempts that students made during the course and their final exam scores. Based on these findings, the answer to the second question is that the generated assessments positively affected students' academic performance, and students who attempted more assessments had a higher tendency to achieve better scores in the final exams of the course than those who attempted fewer ones.

Table 8
Correlations between assessments attempts and Final exam score.

Course	Correlation technique	N	Variable	by Variable	r	Sig.(2-tailed)
Programming 1	Pearson Correlation	120	Assessments Attempts	Final exam score	0.67	< 0.001
	Spearman Correlation	120	Assessments Attempts	Final exam score	0.67	< 0.001
Big Data	Pearson Correlation	70	Assessments Attempts	Final exam score	0.65	< 0.001
	Spearman Correlation	70	Assessments Attempts	Final exam score	0.64	< 0.001
Enterprise Modelling	Pearson Correlation	40	Assessments Attempts	Final exam score	0.59	< 0.001
	Spearman Correlation	40	Assessments Attempts	Final exam score	0.59	< 0.001

5.2. Implications

The findings of this research have significant implications for the development of QA generation tools, especially in the higher educational context. Firstly, it provides valuable insights into the strengths and limitations of different approaches to generating QA pairs, which have proven to be useful in the higher educational context. The integration of these tools into diverse curricula across disciplines introduces a transformative potential for teaching and learning practices. Teachers must be supported through professional development programs to adeptly incorporate these technologies into their strategies.

Secondly, automated QA pairs can enhance student learning outcomes and progress in courses. These tools can revolutionize assessment practices by enabling more formative assessments that provide immediate feedback and support personalized learning paths. Moreover, they encourage higher student engagement and motivation by promoting self-directed learning and critical thinking. Thirdly, the use of automatically generated QA pairs can alleviate the burden on teachers of manually creating multiplechoice questions (MCQs) for their students. This shift can lead to a more efficient allocation of teacher efforts towards more impactful educational activities such as personalized student guidance and curriculum development.

The ethical and societal implications also demand careful consideration. Ensuring equitable access to these tools across different student populations is critical. This involves addressing not only the digital divide but also designing these tools to be inclusive and accessible to students with disabilities. Moreover, the quality of the generated QA pairs is crucial; biases or limitations in the automated generation process must be addressed to ensure the quizzes are unbiased, relevant, and aligned with educational objectives. Institutional strategies and policies will also need to evolve to support the ethical use of these technologies. This includes developing guidelines that address data privacy, combat academic dishonesty, and ensure that these tools align with broader educational goals and standards.

Finally, there is a need for ongoing research into the long-term impacts of these tools on educational outcomes, scalability in diverse educational settings, and the user experience of both students and educators. By expanding the scope of these discussions, the adaptation of automated QA pair generation can contribute positively to the educational context, supporting more efficient and effective learning and assessment practices.

5.3. Limitations and future research

The benchmark datasets utilised in this study were derived exclusively from course materials related to computer science-related subjects. It is important to acknowledge that this strategy may introduce certain limitations in terms of the generalizability of the findings and could potentially introduce bias into the evaluation outcomes. Furthermore, it is important to note that our study focused exclusively on the evaluation of three prominent language models. We did not delve into the investigation of alternative models or the potential synergistic effects of combining different models, which could potentially lead to enhanced performance. The aforementioned limitations underscore the necessity for additional investigation employing larger and more diverse samples, incorporating multiple stakeholders, employing more varied and realistic datasets, and exploring a broader array of models and methodologies. In addition, it is imperative to generate a diverse range of benchmark datasets across various subject areas in order to thoroughly evaluate the efficacy of the chosen methodologies.

6. Conclusion

In the context of higher education, it is the prerogative of the teacher to develop assessments that serve as valuable tools in aiding students in their understanding of various course concepts. In order to conduct thorough assessments, teachers require a collection of QA pairs that are directly relevant to the respective course being taught. In the current study, a deliberate choice was made to employ three distinct methodologies in order to generate automatic QA pairs. These methodologies involved the utilisation of pre-trained large language models, with the primary objective of enhancing educational practises. In order to assess the efficacy of the aforementioned approaches, a comprehensive evaluation was carried out using both automatic evaluation techniques and real-world educational evaluation. The objective was to compare and evaluate the performance of these approaches within the context of educational purposes. The evaluation was conducted on benchmark datasets related to three distinct courses, namely Programming 1, Big Data, and Enterprise Modelling. The findings of this study indicate that the Multi-task approach exhibited superior performance compared to the other two approaches across all benchmark datasets during the automatic evaluation process. The approach utilised for generating QA pairs demonstrated a favourable utility impact, effectively providing support to students. However, there is room for improvement in order to enhance the facilitation of course support. In order to enhance the comprehensiveness and representativeness of the findings, it is recommended that future investigations consider expanding the pool of participants and incorporating a wider range of courses.

CRediT authorship contribution statement

Jintao Ling: Writing – original draft, Software, Methodology, Formal analysis, Conceptualization. **Muhammad Afzaal:** Writing – review & editing, Validation, Supervision, Resources, Methodology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Standards of creating multiple-choice questions

The question creation criteria.

1. Clearly identify the specific point being tested by the item
2. Include the main idea and phrasing in the stem
3. Avoid including irrelevant clues that may lead to the correct answer
4. Eliminate unnecessary or extraneous information in the question stem

5. Use negative words in the stem sparingly, and if used, underline or capitalize them to avoid confusion
6. Ensure that the question is clear and students understand what is expected of them

The options creation criteria.

1. Make sure each option is mutually exclusive and does not overlap with others
2. Ensure that the options are as similar as possible in terms of structure and phrasing
3. Keep the grammar of each option consistent with that of the question stem
4. Ensure that there is only one correct or best answer among the options
5. Use plausible and realistic distractors to challenge the students
6. Incorporate common errors or misconceptions made by students in the distractors
7. Keep the structure of the options parallel and consistent
8. Ensure that the length of the options is similar to avoid giving clues to the correct answer
9. Avoid using the options ”all of the above” and ”none of the above” unless necessary
10. Avoid using absolute determiners such as ”never” and ”always”
11. Randomize the position of the correct answer among the options
12. Use letters instead of numbers to label the options 13. Replace any distractors that are not chosen by any examinees
14. Avoid using humor in the question or options.

References

Akyon, F. C., Cavusoglu, D., Cengiz, C., Altinuc, S. O., & Temizel, A. (2022). Automated question generation and question answering from Turkish texts. https://doi.org/10.48550/arXiv.2111.06476. Comment: 14 pages, 1 figure, 13 tables.

Alberti, C., Andor, D., Pitler, E., Devlin, J., & Collins, M. (2019). Synthetic QA corpora generation with roundtrip consistency. *Proceedings of the 57th annual meeting of the association for computational linguistics*. https://doi.org/10.18653/v1/P19-1620

Chan, Y.-H., & Fan, Y.-C. (2019). A recurrent BERT-based model for question generation. *Proceedings of the 2nd workshop on machine reading for question answering*. https://doi.org/10.18653/v1/D19-5821

Cui, S., Bao, X., Zu, X., Guo, Y., Zhao, Z., Zhang, J., & Chen, H. (2021). OneStop QAMaker: Extract question-answer pairs from text in a one-stop approach [arXiv:2102.12128 [cs]] https://doi.org/10.48550/arXiv.2102.12128.

Gabajiwala, E., Mehta, P., Singh, R., & Koshy, R. (2022). Quiz maker: Automatic quiz generation from text using NLP. In *Futuristic trends in networks and computing technologies* (pp. 523–533). Springer Nature. https://doi.org/10.1007/978981-19-5037-737.

Gao, Y., Bing, L., Li, P., King, I., & Lyu, M. R. (2018). Generating distractors for reading comprehension questions from real examinations. https://doi.org/10.48550/arXiv.1809.02768. Comment: AAAI2019.

Gholami, V., & Morady Moghaddam, M. (2013). The effect of weekly quizzes on students' final achievement score. *International Journal of Modern Education and Computer Science, 5*, 36–41. https://doi.org/10.5815/ijmecs.2013.01.05

Jia, X., Zhou, W., Sun, X., & Wu, Y. (2020). EQG-RACE: ExaminationType question generation. https://doi.org/10.48550/arXiv.2012.06106. Comment: Accepted by AAAI-2021.

Klein, T., & Nabi, M. (2019). *Learning to answer by learning to ask: Getting the best of GPT-2 and BERT worlds*. https://doi.org/10.48550/arXiv.1911.02365 [arXiv:1911.02365 [cs]].

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.

Lelkes, A. D., Tran, V. Q., & Yu, C. (2021). *Quiz-style question generation for news stories*. https://doi.org/10.48550/arXiv.2102.09094 [arXiv:2102.09094 [cs]].

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach [arXiv:1907.11692 [cs]] https://doi.org/10.48550/arXiv.1907.11692.

Mazidi, K., & Nielsen, R. (2014). Linguistic considerations in automatic question generation. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2*, 321–326.

Qi, W., Yan, Y., Gong, Y., Liu, D., Duan, N., Chen, J., Zhang, R., & Zhou, M. (2020). ProphetNet: Predicting future N-gram for sequence-to-sequence pre-training. https://doi.org/10.48550/arXiv.2001.04063. Comment: Accepted to EMNLP 2020 Findings. Project page: github.com/microsoft/ProphetNet.

Qu, F., Jia, X., & Wu, Y. (2021). Asking questions like educational experts: Automatically generating question-answer pairs on real-world examination data. *Comment: Accepted as a long paper in the main conference of EMNLP 2021*. https://doi.org/10.48550/arXiv.2109.05179 [arXiv:2109.05179 [cs]].

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. https://doi.org/10.48550/arXiv.1910.10683. Comment: Final version as published in JMLR).

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. https://doi.org/10.48550/arXiv.1606.05250. Comment: To appear in Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Rodriguez-Torrealba, R., Garcia-Lopez, E., & Garcia-Cabot, A. (2022a). End-to-end generation of multiple-choice questions using text-to-text transfer transformer models. *Expert Systems with Applications, 208*, Article 118258.

Rodriguez-Torrealba, R., Garcia-Lopez, E., & Garcia-Cabot, A. (2022b). End-to-End generation of Multiple-Choice questions using Text-to-Text transfer Transformer models. *Expert Systems with Applications, 208*, Article 118258. https://doi.org/10.1016/j.eswa.2022.118258

Torres, C., Lopes, A. P., Azevedo, J. M., & Babo, L. (2009). Developing multiple choice questions in mathematics [Accepted: 2012-07-30T11:05:13Z]. Retrieved February 19, 2023, from https://recipp.ipp.pt/handle/10400.22/586.

Wang, H., Li, J., Wu, H., Hovy, E., & Sun, Y. (2022). pre-trained language models and their applications. *Engineering*. https://doi.org/10.1016/j.eng.2022.04.024

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.

Yao, B., Wang, D., Wu, T., Zhang, Z., Li, T. J.-J., Yu, M., & Xu, Y. (2022). It is AI's turn to ask humans a question: Question-answer pair generation for children's story books. https://doi.org/10.48550/arXiv.2109.03423. Comment: Accepted to ACL 2022.

Yi, C., Zhu, R., & Wang, Q. (2021). Exploring the interplay between questionanswering systems and communication with instructors in facilitating learning. *Internet Research, 32*(7), 32–55.

Zhong, W., Gao, Y., Ding, N., Qin, Y., Liu, Z., Zhou, M., Wang, J., Yin, J., & Duan, N. (2022). ProQA: Structural prompt-based pre-training for unified question answering [arXiv:2205.04040 [cs]] https://doi.org/10.48550/arXiv.