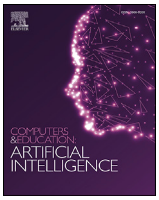
Contents lists available at ScienceDirect

Computers and Education: Artificial Intelligence

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



Analysis of LLMs for educational question classification and generation

Said Al Faraby, Ade Romadhony, Adiwijaya [*]

School of Computing, Telkom University, Jl. Telekomunikasi No.1, Terusan Buah Batu, Bandung, 40257, Indonesia

ARTICLE INFO

Keywords:
Question classification
Question generation
Graesser's taxonomy
LLMs

ABSTRACT

Large language models (LLMs) like ChatGPT have shown promise in generating educational content, including questions. This study evaluates the effectiveness of LLMs in classifying and generating educational-type questions. We assessed ChatGPT's performance using a dataset of 4,959 user-generated questions labeled into ten categories, employing various prompting techniques and aggregating results with a voting method to enhance robustness. Additionally, we evaluated ChatGPT's accuracy in generating type-specific questions from 100 reading sections sourced from five online textbooks, which were manually reviewed by human evaluators. We also generated questions based on learning objectives and compared their quality to those crafted by human experts, with evaluations by experts and crowdsourced participants.

Our findings reveal that ChatGPT achieved a macro-average F1-score of 0.57 in zero-shot classification, improving to 0.70 when combined with a Random Forest classifier using embeddings. The most effective prompting technique was zero-shot with added definitions, while few-shot and few-shot + Chain of Thought approaches underperformed. The voting method enhanced robustness in classification. In generating type-specific questions, ChatGPT's accuracy was lower than anticipated. However, quality differences between ChatGPT-generated and human-generated questions were not statistically significant, indicating ChatGPT's potential for educational content creation. This study underscores the transformative potential of LLMs in educational practices. By effectively classifying and generating high-quality educational questions, LLMs can reduce the workload on educators and enable personalized learning experiences.

1. Introduction

Questions occupy a central and indispensable role within the educational landscape, serving as fundamental tools for fostering comprehension and knowledge acquisition (Chin & Osborne, 2008). These questions cover a range of difficulty and usefulness, with some being better for learning than others. The ability to craft valuable questions constitutes a skill that necessitates not only linguistic proficiency but also a profound understanding of pedagogical objectives (Nappi, 2017). This process requires significant mental effort and a deep understanding of how learning works.

In response to the complexities associated with question formulation, the proposition of a system capable of not only analyzing questions but also automatically generating them emerges as a valuable resource within the educational domain. Such a system has the potential to alleviate the burden on educators and learners alike by facilitating the creation of well-structured and pedagogically effective questions.

Historically, Automatic Question Generation (AQG) relied heavily on rule-based methods. These approaches used hand-crafted linguistic rules and resources, such as Part-of-Speech (POS) tags and syntax labels, to transform declarative sentences into questions (Haris & Omar, 2012; Heilman & Smith, 2010). While these methods could generate questions that were syntactically sound, they required extensive manual effort and linguistic expertise, limiting their scalability and adaptability across diverse educational content.

The advent of neural-based methods marked a significant shift in AQG (Zulqarnain et al., 2021; Du et al., 2017). These methods utilize large datasets and advanced machine learning algorithms to identify patterns and generate questions more effectively than rule-based approaches. Despite these advancements, neural models still encounter several challenges. They often struggle with generating questions from long reading passages and abstract answers. Additionally, many neural models have been primarily trained on low-level questions, limiting their ability to create more complex and cognitively demanding questions (Al Faraby et al., 2023).

Within the broader context of Natural Language Processing (NLP), the emergence of LLMs, which is the latest advancement of neural-based

[*] Corresponding author.
E-mail address: adiwijaya@telkomuniversity.ac.id (Adiwijaya).

https://doi.org/10.1016/j.caeai.2024.100298
Received 23 January 2024; Received in revised form 19 August 2024; Accepted 9 September 2024

Available online 12 September 2024

methods, has introduced a paradigm shift. These models, characterized by their extensive pretraining on vast textual corpora, have garnered widespread attention and utilization across various NLP applications. Some of the well-known instances include Generative Pre-trained Transformer (GPT) (Brown et al., 2020), Large Language Model Meta AI (LLaMA) (Touvron et al., 2023), and Pathways Language Model (PaLM) (Chowdhery et al., 2022).

The rationale for selecting LLMs for this research lies in their demonstrated capability to understand and generate human-like text across a wide array of contexts including long context, making them ideal candidates for educational applications. Given their ubiquity and versatility, it is becoming crucial to study how well they work when used for educational questions. This study is significant in today's NLP research because it aims to understand how these models can be beneficial in education.

Despite these advancements, there are specific gaps in current research. While LLMs can produce questions that are syntactically and semantically superior, the quality and pedagogical effectiveness of these questions have not been thoroughly evaluated. Moreover, the ability of LLMs to generate questions that align with educational objectives and stimulate critical thinking remains underexplored.

Question classification is another critical aspect of enhancing educational tools. Effective classification can help organize and assess educational content by categorizing questions based on various types, such as Bloom's taxonomy (Bloom, 1956) and Graesser's question typologies (Graesser & Person, 1994). This capability is crucial not only for improving the variation and usability of questions but also for ensuring precision after question generation.

Automatic question classification and AQG can significantly improve teaching and learning environment. Those systems can be seamlessly integrated into websites or reading software, transforming the reading experience from passive to interactive by generating questions on-the-fly based on the current page, thus prompting readers to engage actively with the material (Syed et al., 2020). Additionally, the systems allow for the customization of questions tailored to individual learning needs, enabling learners to focus on areas that need improvement by receiving questions that match their proficiency level (Srivastava & Goodman, 2021). Furthermore, well-designed automated questions can promote critical thinking by encouraging students to analyze information and apply knowledge to real-world scenarios. Beyond answering the automated questions, students can also learn to formulate critical questions themselves by observing the examples provided, thereby enhancing their ability to ask more insightful and effective questions (Hofstein et al., 2005).

Given these factors, this study aims to conduct a thorough evaluation of how well large language models perform in the context of classifying and generating educational questions. Through empirical evaluation and analysis, this research aims to understand the strengths and limitations of these models in the field of educational questioning. Ultimately, the goal is to provide information and guidance on how to use these models effectively in educational settings.

Specifically, the research questions addressed in this paper are as follows:

- *RQ1: How effectively do LLMs perform in classifying educational-type questions?* Understanding the classification performance of LLMs in educational contexts is essential for assessing their applicability and reliability in sorting and organizing educational content. This knowledge is crucial for developing systems that can automatically categorize questions and potentially enhance automatic question generation systems.
- *RQ2: How accurate are LLMs in generating type-specific questions?* Evaluating the accuracy of LLMs in generating specific types of questions helps in assessing their capability to produce relevant and targeted educational content. This is important for ensuring that generated questions are useful and appropriate for different educational scenarios, enhancing the learning experience.
- *RQ3: How does the quality of questions generated by LLMs compare to those created by human experts?* Comparing the quality of LLM-generated questions to those created by human experts provides insights into the effectiveness of LLMs in producing high-quality educational content. This helps in identifying areas where LLMs excel or need improvement, guiding future enhancements and applications in educational settings.

2. Related work

The related work in this paper covers various aspects of automatic question generation, educational question classification, and the applications of large language models (LLMs) in educational contexts. These interconnected topics collectively inform the current state of research and highlight the advancements and challenges in leveraging AI for educational purposes.

2.1. Automatic question generation (AQG)

Automatic Question Generation (AQG) is the process of generating syntactically fluent and semantically relevant questions from an input context. Historically, research in question generation relied heavily on rule-based methods. These approaches utilized hand-written rules, linguistic features (such as Part-of-Speech (POS) tags, entities, and syntax labels), and language resources (e.g., WordNet) that necessitated a deep understanding of linguistics (Ali et al., 2010; Mitkov & Le An, 2003; Heilman & Smith, 2010; Mostow & Chen, 2009). Rule-based AQG methods often focused on transforming declarative sentences into questions using syntactic patterns and semantic constraints. For example, Heilman and Smith (2010) developed a system that used handcrafted rules to generate questions from declarative sentences, emphasizing the importance of syntactic transformation rules and lexical resources. While effective to some extent, these methods required extensive manual effort and expertise, limiting their scalability and adaptability to diverse educational content.

The advent of neural-based methods marked a significant shift in AQG. These approaches leverage large datasets and machine learning algorithms to learn patterns and generate questions more effectively. Du et al. (2017) introduced an end-to-end, sequence-to-sequence (seq2seq) system for AQG, which utilized recurrent neural networks (RNNs) with attention mechanisms. Trained on the SQuAD dataset (Rajpurkar et al., 2016), this model demonstrated significant improvements over rule-based systems, achieving a BLEU4 score of 12.28 compared to 11.18 achieved by Heilman and Smith (2010).

RNN-based seq2seq models, however, face challenges related to computational cost and handling long-distance dependencies. The introduction of transformers and pre-training on extensive text datasets has addressed these issues. For instance, a recurrent BERT-based model (Chan & Fan, 2019) achieved a significant improvement in BLEU4 score, reaching 22.17. Furthermore, a specialized pre-training and fine-tuning framework for transformer-based text generation models attained an even higher BLEU4 score of 26.95 (Xiao et al., 2020). These advancements leveraged advanced pre-training and fine-tuning techniques to enhance the natural language generation process, resulting in questions that are both syntactically and semantically superior.

Despite these advancements, evaluating the quality of generated questions remains challenging. Traditional automatic metrics such as BLEU and ROUGE often fail to capture the nuances of question quality, necessitating human evaluations for a more comprehensive assessment (Mathur et al., 2020; Sultan et al., 2020). Due to the limitations of automated metrics, many studies also include manual evaluations by humans, focusing on criteria such as naturalness (fluency, relevance, and answerability), difficulty (Du et al., 2017; Chan & Fan, 2019; Bi et al., 2021), and helpfulness (Cheng et al., 2021; Sekulić et al., 2021).

However, the lack of standardized evaluation methods makes it difficult to compare results across studies. Evaluations typically involve experts, crowdsourced participants, or the authors, who assess the questions based on specific criteria and assign scores. Some studies also compare preferences between human-generated and AI-generated questions. For educational purposes, Horbach et al. (2020) proposed a comprehensive human evaluation scheme that assesses the quality of generated questions with nine criteria, considering their relevance, complexity, and importance to the educational context. This tiered evaluation approach highlights the value of questions created by domain experts, ensuring their pedagogical effectiveness in educational settings.

2.2. Educational question classification

Building on the theme of improving educational tools, question classification plays a crucial role in organizing and assessing educational content. One of the foundational studies on question classification in education by Fei et al. (2003) aimed to classify multiple-choice questions into three levels of difficulty: easy, medium, and hard. They utilized neural networks along with linguistic features such as term frequency and the length of the questions and answers, achieving an impressive F1-score of 78%. Meanwhile, other researchers have focused on categorizing questions based on Bloom's taxonomy. For example, Haris and Omar (2012) used a rule-based classifier on a small dataset of 135 questions, achieving an F1-score of 77%. Similarly, Yahya and Osman (2011) employed TF-IDF features and SVM classifiers on 190 questions across six categories, reaching an accuracy of 87.4%, though with a lower F1-score of 44.64% due to recall issues. More recently, Mohammed and Omar (2020) applied TFPOS-IDF and pre-trained word2vec as feature extractors, and experimented with KNN, Linear Regression, and SVM classifiers. The SVM classifier yielded the highest performance, with a weighted F1-score of 89.7%.

In addition to Bloom's taxonomy, question classification has also been explored using Graesser's typology. Cao and Wang (2021) collected user-generated questions from online forums and labeled 5,000 data points with 10 question types derived from Graesser's typology. By utilizing a pre-trained RoBERTa model, they achieved a macro F1-score of 0.80. This work highlights the effectiveness of advanced pre-training models in handling diverse question types and improving classification performance.

2.3. LLMs applications in education

Shifting focus to the broader applications of LLMs in education, these models have revolutionized the field with their vast capabilities. LLMs are an evolution of Pre-trained Language Models (PLMs) with much larger model sizes, significantly more data, and much longer training times. In addition to improved performance on downstream tasks, LLMs demonstrate novel capabilities (emergent abilities) such as in-context learning, instruction following, and step-by-step reasoning (Zhao et al., 2023). With these new abilities, their usage has become widespread, particularly in the field of education.

For example, ChatGPT, one of the most well-known LLMs, has been shown to provide valid insights on topics related to clinical education that can aid in learning (Kung et al., 2023). Zhai (2023) demonstrated that ChatGPT is capable of making assessments that meet certain performance expectations. Abdelghani et al. (2022) employed GPT-3 to generate cues to enhance primary school children's question-asking skills. These studies underscore the potential of LLMs to support various educational tasks, from providing explanations and feedback to generating educational content.

2.4. LLMs in educational question classification and generation

Bridging the discussion on AQG and educational question classification, recent studies have evaluated LLMs for various educational tasks, highlighting both their capabilities and limitations. For instance, Koto et al. (2023) evaluated LLMs' performance in answering questions across 64 different tasks and education levels, finding that while LLMs like ChatGPT performed adequately at the elementary level, their performance varied across tasks and levels. Meanwhile, Kasneci et al. (2023); Crompton and Burke (2024) discussed the opportunities and challenges of using ChatGPT for educational purposes, including question generation. However, these discussions lacked deep experimental analysis on the quality and effectiveness of the generated questions.

Exploring the potential of ChatGPT for question generation, Cooper (2023) engaged ChatGPT in a dialogue and tasked it with generating multiple-choice questions based on a given topic. Despite the relevance of their approach, the study lacked rigorous experimental analysis. Similarly, Xiao et al. (2023) focused on generating passage and multiple-choice questions for reading comprehension, but their generated questions were criticized for exhibiting obvious patterns, being overly straightforward, and lacking variation. Lastly, Olney (2023) concentrated on generating multiple-choice questions from textbooks and comparing them with human-generated questions, providing insights into the quality of questions produced by ChatGPT but without a thorough experimental setup. These studies collectively highlight the potential of LLMs for educational question generation but also underscore the need for deeper analysis and experimentation to improve the quality and effectiveness of the generated questions.

To our knowledge, there are no specific studies focusing on the use of LLMs for question classification. However, for general text classification, Sun et al. (2023) demonstrated that LLMs could achieve state-of-the-art performance on several benchmark datasets. This suggests that while LLMs have not been explicitly studied for question classification, their demonstrated capabilities in text classification indicate a promising potential for this application. Further research is needed to explore and validate the effectiveness of LLMs in classifying educational questions, which could significantly enhance the efficiency and accuracy of educational assessments.

3. Methods

This section explicitly links each research question to the specific characteristics extracted, the methodologies used, and how the analysis was conducted.

3.1. RQ1: effectiveness of LLMs in classifying educational-type questions

Extracted Characteristics: To assess how effectively LLMs perform in classifying educational-type questions, we extracted several characteristics:

1. *Classification Performance (F1-Score)*: We measured the overall performance of the LLMs using the Macro-average F1-Score, which combines precision and recall into a single metric reflecting both accuracy and completeness in classification. A high F1-Score indicates the LLM's effectiveness in accurately classifying educational-type questions, providing a balanced measure of its performance.
2. *Impact of Prompting Techniques*: We compared the effectiveness of different prompting techniques to determine their influence on classification accuracy. The prompting techniques include:
   - **A basic zero-shot prompt** containing only classification instructions and a list of question types.
   - **An enhanced zero-shot prompt** that includes descriptions of each question type in addition to the basic instructions.
   - **A prompt utilizing few-shot learning**, which includes classification instructions, a list of question types, detailed descriptions, and examples of each question type to improve the model's understanding and accuracy.
   - **A few-shot learning prompt with Chain of Thought (CoT)**, similar to the previous prompt but with the addition of the CoT

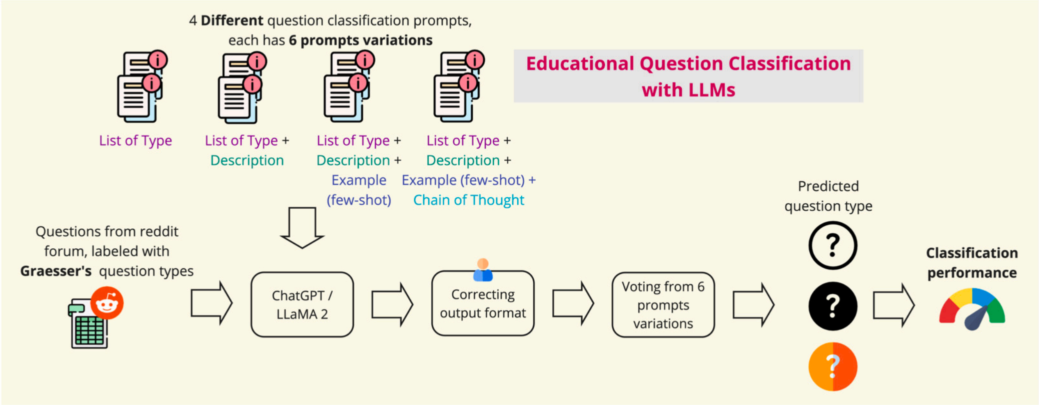

4 Different question classification prompts, each has 6 prompts variations
Educational Question Classification with LLMs
List of Type
List of Type + Description
List of Type + Description + Example (few-shot)
List of Type + Description + Example (few-shot) + Chain of Thought
Questions from reddit forum, labeled with Graesser's question types
ChatGPT / LLaMA 2
Correcting output format
Voting from 6 prompts variations
Predicted question type
Classification performance

Fig. 1. Research design for question classification task.

technique to encourage more detailed reasoning and improved performance.

By analyzing the classification performance with each prompt, we can identify the prompting technique that yields the highest accuracy. This analysis helps us understand the impact of additional information, such as descriptions, examples, and the Chain of Thought (CoT) technique, on the model's performance. Ultimately, this reveals the most effective prompting technique for classifying educational-type questions, indicating the optimal way to instruct the LLM for improved performance.

3. *Effectiveness of the Voting Method*: We analyzed the impact of using a Voting method to aggregate results from multiple prompts, aiming to improve classification robustness and accuracy by reducing random effects. The analysis shows whether the Voting method enhances the LLM's classification performance, indicating its effectiveness in improving the accuracy and reliability of the model's predictions.
4. *Accuracy in Classifying Self-Generated Questions*: One scenario to ensure the accuracy of AQG is through post-filtering (Heilman & Smith, 2010). Post-filtering can be achieved by classifying the generated questions. The characteristics of generated questions may differ from user-generated questions, thus the effectiveness of LLMs needs to be tested for these as well. This ensures that the model is robust and accurate in different scenarios, validating its utility in AQG systems.

Methodology: The methodology to answer the first research question is as follows, as illustrated in Fig. 1:

Preparing the Evaluation Dataset: The dataset used in this question classification task consists of user-generated questions labeled into ten categories adapted from Graesser's taxonomy, as provided by Cao and Wang (2021). Out of the total 4595 questions and labels, we performed an 80:20 split using a stratified method to ensure consistent proportions of each label. The 20% subset was used for evaluating the classification performance through prompting. The remaining 80% served as training data for other classification techniques, which were compared to the prompting techniques. The examples used in the third (few-shot) and fourth (few-shot & CoT) prompts were also drawn from this split.

The label statistics for this dataset reveal a Krippendorff's α of 0.67, indicating a moderate level of agreement among annotators for all samples (Marzi et al., 2024). In the context of testing data, the agreement between two annotators on their first choice stands at 69.7% (691 out of 992). Furthermore, considering agreement on any choice, the two annotators exhibit a higher consensus rate of 77.6% (770 out of 992). These figures underscore the reliability and consistency of the labeling process.

However, it is important to note that label agreement is not perfect, indicating potential bias due to annotator interpretation or the inherent ambiguity of some questions. Additionally, we acknowledge a limitation in the dataset's representativeness. Since the data was sourced from online question-and-answer forums, it may not fully capture the broad spectrum of educational questions, particularly those common in formal educational settings.

Prompting and Classification: Research shows that LLMs are not robust against prompt variation (Mizrahi et al., 2023; Gu et al., 2023). The performance of LLMs varies widely with small changes to the prompt, such as paraphrasing, even though they are semantically equivalent. To improve LLMs' robustness in this question classification task, we created six variations of prompts for each of the four prompt techniques, resulting in a total of 24 prompts. These variations differ in the wording and structure of the instructions. This approach is based on concepts of self-consistency by sampling multiple outputs (Wang et al., 2022) and DiVeRSE techniques that diversify the prompt and use voting to enhance performance (Li et al., 2023). Both self-consistency and DiVeRSE generate multiple reasoning paths that are combined to form a final answer.

Each question in the evaluation dataset was then classified using each of these prompts. We have chosen the OpenAI-provided LLM model for our main model, owing to its superior performance, as well as its accessible API allowing us to deploy models without manual intervention. For this task, we utilized the API's default parameters, following manual exploration that yielded satisfactory results.

Correcting Output: In the context of prompt-based classification, it is often necessary to perform post-processing to arrive at the definitive answer. Responses generated by LLMs may exhibit various characteristics that necessitate further refinement. These characteristics may include the presence of words not explicitly associated with the designated label, variations in word forms compared to the label's specifications, the provision of multiple answers, or the possibility that ChatGPT does not categorize the response within any predefined label.

The post-processing involved several stages: lowercasing, removing non-alphanumeric characters, and replacing certain frequent non-label words with predefined labels, such as 'none of the given' → 'other', 'procedure' → 'procedural', and 'causal' → 'cause', based on statistical observations of LLM outputs. After these steps, string matching was performed to identify predefined labels. If a label could not be determined through these processes, the output was categorized as 'other'.

The most common error was that LLMs sometimes expanded the output when only letter options were expected, for example, turning 'a', 'c' into 'a. verification', 'c. concept'. Conversely, LLMs sometimes reduced the output unnecessarily. A manual analysis revealed that only about 0.2% of ChatGPT outputs were non-label words that could be associated with label words due to typos, different word forms, or the use of similar words. Furthermore, the total 'other' labels amounted to 155 out of 23,808 outputs (derived from 992 questions x 4 prompts x 6 prompt variations), representing just 0.6% of the total. Notably, most

Table 1
An illustration of a question generation prompt that produces multiple questions from a single reading section.

Prompt	Output
Causal Antecedent Question: Asking for state or event causally led to an event or state. Instruction: Based on the reading section below, create some "causal antecedent questions" and use the following format. Q1: Q2: Qn: Here is the reading section: "' ⟨ Sample reading section ⟩ "'	Q1: What economic entities are the suppliers of loanable funds in the market? Q2: How does a higher interest rate affect the quantity supplied of loanable funds? Q3: Who are the demanders of loanable funds in the market? Q4: How does a higher interest rate impact the quantity demanded of loanable funds? Q5: What determines the equilibrium interest rate in the market for loanable funds?

of these 'other' labels (117 or 75%) occurred in the second variation of prompt-4, indicating that ChatGPT was unable to classify the question into one of the predefined labels. This issue might be a negative side effect of the multiple steps in the Chain of Thought (CoT) process, which ultimately caused confusion for ChatGPT. Overall, these statistics indicate that ChatGPT's outputs generally do not require complex post-processing.

Aggregating Classification with Voting: To aggregate the outputs from all six variations of prompts, as mentioned in step 2, we employed a voting method. The voting process involved the following steps:

- Collecting Outputs: Gather classification outputs from each prompt for a given question.
- Determining Consensus: Identify the most frequently occurring classification among the outputs. If there is a tie, the label corresponding to the first occurrence is chosen.
- Final Classification. Assign the classification determined by the consensus as the final label.

Performance Analysis: We compared the model's classifications with the human labels available in the evaluation dataset. We then calculated the F1-score for each class. To obtain the overall performance, we aggregated the F1-scores of all classes using the macro-average method. It is important to note that for this multi-category classification task, the F1-Score is calculated for each category individually and then averaged using the macro-average method to provide a comprehensive performance metric across all categories. We chose the macro-average method because the data proportions for each class varied significantly, ranging from as low as 3% to as high as 20%, ensuring that the overall performance was not dominated by the majority class. We calculated performance for each prompt variation as well as the final prompt from the Voting method.

3.2. RQ2: accuracy of LLMs in generating type-specific questions

Extracted Characteristics: To evaluate the accuracy of LLMs in generating type-specific questions, we analyze the following key characteristics:

1. *Accuracy of Generated Question Type*: This characteristic measures how accurately the generated questions match the specified types in the prompts. It evaluates the model's ability to produce questions that conform to the requested educational categories, such as comparison, procedural, or causal questions. This analysis is crucial for assessing the LLM's performance in generating type-specific questions accurately.
2. *Effect of Generation Sequence on Accuracy*: LLMs can generate a sequence of multiple questions from a single prompt execution, as illustrated in Table 1. This characteristic examines how the order of question generation affects the accuracy of the question types. It analyzes whether the position in the sequence influences the model's ability to produce correct and relevant questions, providing insights into the model's consistency and reliability over successive generations.
3. *Incidence of Hallucinations in Generated Questions*: For the purpose of understanding the material, the generated questions are expected to be based solely on the provided source text. Therefore, if the content of the questions or answers includes information beyond the source text, it is considered undesirable and may be classified as hallucination. This characteristic identifies the occurrence of such hallucinations. By identifying and categorizing these deviations from the expected content, we can assess the reliability of the generated questions. A lower incidence of hallucinations indicates higher reliability and accuracy in question generation.

Methodology: The following processes were undertaken to analyze the key characteristics:

Preparing Educational Text: We collected data from online textbooks provided freely by OpenStax.com. There were 5 textbooks used, which are Principles of Finance, Introduction to Philosophy, Anatomy and Physiology, Biology, and Contemporary Mathematics. 20 sections from each of the textbooks were randomly selected and spread over the different chapters, resulting in a total of 100 sections. During the manual scrapping process, only the main texts were extracted, leaving out the tables, pictures, videos, and additional texts that usually come in special boxes as well as the summary at the end. On average, the number of words per reading section was 1513.08, with the shortest having 317 words and the longest 6190 words. Apart from extracting the text, the learning objectives from each of the sections were also taken. The learning objectives data are then used for one of the experiments in question generation.

Prompting and Generation: We provided the LLM with prompts specifying four different question types, *antecedent*, *consequence*, *comparison*, and *procedural*. The model then generated questions based on these prompts, ensuring a variety of generated question types for evaluation. Here is the detailed configuration of the model, which was determined based on recommendations from the community forum.[1]

- model = "gpt-3.5-turbo-16k-0613"
- temperature = 0.5
- top-n = 0.5
- frequency penalty = 1
- presence penalty = 1

The temperature and top-n were set to 0.5 to ensure a balance between coherence and diversity, while the frequency and presence penalties were set to 1 to help avoid repetition of words and phrases across

[1] https://community.openai.com/t/cheat-sheet-mastering-temperature-and-top-p-in-chatgpt-api/172683.

different questions, as we aimed to generate multiple questions simultaneously. We tracked the order in which each question was generated to evaluate the effect of generation sequence on accuracy.

Manual Evaluation: To evaluate the accuracy of the generated question types, we employed two human annotators, one with a graduate qualification and the other with a postgraduate qualification. From 100 sections of input text, we sampled 50 sections for manual evaluation. For each section, the first and last questions generated were selected, resulting in a total of 100 questions from each prompt and each LLM for manual evaluation. This yielded a grand total of 1000 questions (100 questions x 5 prompts x 2 LLMs) for manual annotation.

Before the actual evaluation, five candidate evaluators underwent a training session where they familiarized themselves with the labeling process and practiced on a question dataset. Afterward, we manually reviewed their labeling results and selected two final evaluators.

The two evaluators manually reviewed and labeled the generated questions, assigning each question to one of five labels: *antecedent*, *consequence*, *comparison*, *procedural*, or *other*. The annotators were provided with detailed labeling guidelines, including descriptions and examples for each question type (Appendix A.2). These guidelines were intended to standardize the evaluation process and reduce subjective bias. The evaluators were not informed of the requested type for each generated question to avoid bias, ensuring that the assessment of the model's performance was as objective as possible.

The labeling process was conducted in two stages. In the first stage, each evaluator independently assigned labels to each question. In the second stage, they discussed any discrepancies in their labels to reach a consensus. This second stage followed the labeling process outlined by Cao and Wang (2021).

Source-Constrained Question Generation Analysis: This investigation aims to determine if the questions generated using the text as the source incorporate content from outside the source text. The idea is to identify counterexamples where the generated questions contain information not present in the source text. To achieve this, we provide source texts that do not include comparison or procedural information and then ask LLMs to generate questions of these types. The resulting questions are analyzed to see if LLMs can produce the specified question types without adding external information.

The experiment consists of the following steps:

1. **Dataset Creation**: Create a small dataset of 10 definitional sentences, generated by LLMs (GPT-4) and manually checked for accuracy.
2. **Question Generation**: Ask LLMs to generate comparison and procedural questions, testing different variations such as using GPT-3.5 vs GPT-4 and prompts that generate only questions or both questions and answer snippets.
3. **Evaluation**: Review each generated question and answer to determine if the question type is correct (CT) or incorrect (WT), and if the answer snippet is appropriate (CA) or incorrect (WA). Additionally, mark responses as “Not Compatible” (NComp) if the requested question types cannot be created from the source text. By identifying and categorizing these counterexamples, we can assess the reliability of the generated questions.

Performance Analysis: We calculated the accuracy as the proportion of correctly generated questions out of the total generated questions for each requested type. This provided a clear measure of the model's ability to generate the requested question types accurately. We analyzed the correctness of the questions in relation to their order of generation. Statistical methods were used to identify any trends or patterns indicating how the position in the generation sequence affected correctness.

3.3. RQ3: quality comparison between LLM-generated and human-generated questions

Extracted Characteristics: These are the primary aspects that will be examined and analyzed to compare the quality between questions generated by Large Language Models (LLMs) and those generated by humans:

1. *Quality Comparison Based on Expert Evaluators*:
   This characteristic focuses on the evaluations provided by subject matter experts. These experts possess in-depth knowledge of the subject matter in the questionnaire and have experience in creating educational questions. Their expertise ensures a comprehensive understanding of educational goals and the nuances of question quality. The evaluation criteria used by expert evaluators include:
   - Clarity: Ensures that questions are easy to understand, minimizing student confusion. This criterion, along with the next three, is adapted from the criteria proposed by Horbach et al. (2020).
   - Alignment with Learning Objectives: Checks that questions are relevant to the educational goals, supporting targeted learning outcomes.
   - Stimulation of Critical Thinking: Measures the questions' ability to promote deeper cognitive engagement.
   - Overall Usefulness: Gauges the practical value of the questions in an educational context.
   - Difficulty Level: Ensures that questions are appropriately challenging for the intended audience, facilitating effective assessment. This criterion provides a simpler measure for non-experts in education compared to the critical thinking criterion and has been used in previous studies (Gao et al., 2019; Kumar et al., 2019).
   - Resemblance to Human-Like Questioning: Assesses whether AI-generated questions match the quality and style of human-generated questions. This criterion is inspired by the Turing Test and has been used in prior research (Nov et al., 2023).
2. *Quality Comparison Based on Crowdsourced Evaluators*: This characteristic examines the evaluations provided by a diverse group of crowdsourced evaluators from Amazon Mechanical Turk. The evaluators meeting the following criteria: more than 50 tasks approved and a task approval rate above 98%, located in the United States as a proxy for native English speakers, and having a job function in education and training. The evaluation criteria used by crowdsourced evaluators include the same six categories.

By comparing evaluations from human experts and crowdsourced evaluators, we gain insights into different perspectives of question quality. While experts bring depth and subject-specific insights, crowdsourced evaluators offer diverse perspectives that can reveal broader usability and accessibility issues. This comprehensive evaluation approach provides a well-rounded assessment of the quality of LLM-generated questions compared to those crafted by human experts.

Methodology. To comprehensively evaluate the quality of questions generated by ChatGPT compared to those created by human experts, we employed a structured methodology involving the preparation of comparative question sets and an A/B evaluation by two groups of evaluators:

Preparation of Comparative Question Sets: The purpose of this investigation is to compare the quality of questions generated by ChatGPT with those from textbooks, which are assumed to be created by human experts. From the five textbooks sourced from openstax.org, only the Philosophy textbook included essay questions linked to specific sections. This ensures that the questions are clearly sourced from specific material, which can then be used to generate corresponding questions using ChatGPT. Other textbooks had review questions at the end of chapters, but these were not linked to specific sections, making it difficult to identify the exact source material.

Table 2
Classification Performance on User-Generated Questions (Macro-average F1-Score).

Prompt Variation	Prompt-1 F1-Score	Prompt-2 F1-Score	Prompt-3 F1-Score	Prompt-4 F1-Score
1	0.43	**0.55**	**0.55**	0.45
2	0.43	0.52	0.48	0.37
3	0.44	0.48	0.47	0.43
4	0.39	0.52	**0.55**	**0.49**
5	0.42	0.51	0.49	0.45
6	**0.50**	0.45	0.39	0.39
Average	0.44	0.51	0.49	0.43
Std	0.0362	0.0351	0.0595	0.0438
Voting	**0.52**[a]	**0.57**[a]	**0.56**[a]	**0.52**[a]

The “Voting” row indicates the F1-Score achieved by aggregating the results from multiple prompt variations using a voting method, as discussed in the text.
[a] The voting F1-score outperforms the scores of any individual prompt variation.

We selected five reading sections from the Philosophy textbook along with their learning objectives and a few review questions from each section. Using these five selected reading sections and their learning objectives, we prompted ChatGPT to generate several questions based on this material. This process resulted in five sets of comparison questions generated by ChatGPT. Both sets of questions (from the textbook and ChatGPT) will then be provided to evaluators in the subsequent evaluation stage.

A/B Evaluation of Questions: After obtaining sets of questions from both human experts and ChatGPT for the five reading sections, we created five evaluation forms. The information provided on each evaluation form includes the learning objectives, Set of Question A, and Set of Question B. The assignment of human-generated questions and ChatGPT-generated questions to Set A or Set B is random. This means that in some evaluation forms, Set A contains human-generated questions, while in others, Set A contains ChatGPT-generated questions. The evaluators were not informed about the source of the questions in each set, nor were they told that the comparison involved questions from humans and ChatGPT. This approach was adopted to encourage evaluators to provide assessments that were as unbiased and natural as possible, without attempting to discern which set of questions originated from a human expert.

Next, evaluators were asked to choose their preference for each quality criterion described above. They had three options: “A > B” (indicating that Set A is superior to Set B in terms of clarity, for instance), “B > A”, or “Cannot decide”. An illustration of the evaluation form is shown in Fig. A.5.

We then provided the evaluation forms to two groups of evaluators. The first group, referred to as Experts, consisted of five PhD students who had previously taken a philosophy course. They all also have teaching experience in higher education institutions, with two evaluators under 30 years old, two between 30 and 40 years old, and one over 40 years old. The group included two female evaluators and three male evaluators. The second group was a mix of educators and students with backgrounds in mathematics, economics, and IT. Their familiarity with the subject matter varied: some had taught the philosophy subject presented in the questionnaire, others had studied it, and some had only heard of it without formal education.

As this study involved human, their privacy rights were strictly observed. The participants were protected by hiding their personal information during the research process. They knew that the participation was voluntary and they could withdraw from the study at any time. There is no potential conflict of interest in this study. The data can be obtained by sending request e-mails to the corresponding author.

Performance Analysis: The performance analysis involved calculating the proportion of evaluator preferences for each criterion toward human-generated or ChatGPT-generated questions. To determine whether the differences in preferences for each criterion were statistically significant, we conducted a series of binomial tests. These tests compared the proportions of preferences, excluding the ‘cannot decide’ responses, to assess the statistical significance of any observed differences.

The null hypothesis (H_o) for these tests posits that the proportion of evaluators preferring human-generated questions is equal to the proportion preferring ChatGPT-generated questions, implying that any observed difference in preferences is due to random chance, with no inherent preference for either type of question. The results of the binomial tests would indicate whether the differences in preferences for each criterion are statistically significant, thereby providing a rigorous and unbiased assessment of the comparative quality of questions generated by ChatGPT and human experts.

4. Results

This section presents the findings of our study on the effectiveness of large language models (LLMs) in educational question classification and generation. We evaluate the performance of two models, analyze their accuracy in generating type-specific questions, and compare the quality of questions generated by LLMs to those created by human experts. The results provide insights into the capabilities and limitations of LLMs in educational contexts, addressing the research questions outlined earlier.

4.1. RQ1: effectiveness of LLMs in classifying educational-type questions

Table 2 shows the classification performance of questions using four types of prompts, each with six variations. The performance metric used here is the F1-Score, which is a measure of a test’s accuracy. It considers both the precision and the recall of the test to compute the score. The average F1-Score of each prompt technique is presented for comparison with the Voting method.

4.1.1. Classification performance

It is important to note that the average F1-Score in Table 2 represents the overall expected performance across different variations. This approach accounts for the variability and reliability of each model’s performance, providing a more accurate reflection of the model’s overall capability. The highest expected performance was achieved with the Prompt-2 technique, which included a list of question types along with a description of each question type.

Table 3 provides a more detailed analysis of the performance for each question type. Among the top three types, which are cause, procedural, and extent, all of them have F1-scores above 0.7. On the other hand, the bottom three types, example, ‘verification’, and ‘consequence’, all have F1-scores below 0.5. Specifically for ‘example’ and ‘verification’, the issue lies in their very low recall. Looking at the confusion matrix in Fig. 2, it can be observed that these two types are often mispredicted as ‘judgmental’. This aligns with the analysis presented by Faraby et al. (2022), which suggests the presence of implicit ‘judgmental’ elements in questions that are actually of different types, leading to ambiguity for both human annotators and prediction models.

Table 3
Classification Results per Question Type for Prompt-2 (Voting).

Class	precision	recall	f1-score	support
verification	0.62	0.33	0.43	112
disjunction	0.39	0.59	0.47	49
concept	0.55	0.55	0.55	83
extent	0.59	0.93	0.72	67
example	0.83	0.28	0.42	208
comparison	0.41	0.68	0.51	41
cause	0.82	0.9	**0.86**	133
consequence	0.39	0.53	0.45	30
procedural	0.71	0.81	0.76	135
judgmental	0.43	0.6	0.5	134
macro avg	0.57	0.62	0.57	992

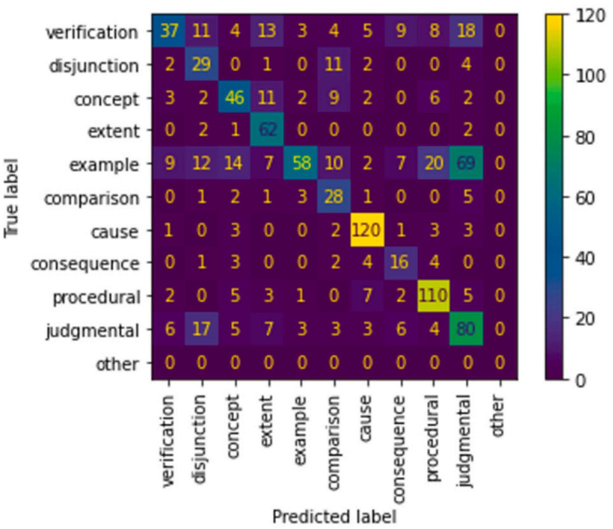


Fig. 2. Confusion matrix comparing Human label to ChatGPT's prediction using the Prompt-2.

Comparing the performance of one of the most powerful LLMs with other machine learning models on the same task provides valuable insights. Table 4 shows the results of using embeddings followed by a classifier and also fine-tuning a pretrained model. The Roberta model by Faraby et al. (2022) achieves an F1-score of 0.81, significantly higher than ChatGPT's score of 0.57. While both models were evaluated on the same dataset, it is important to note that the exact splits used for validation/testing are not identical. However, this substantial difference in performance highlights the potential advantages of fine-tuning for specific tasks.

On the other hand, prompting ChatGPT has a higher F1-score compared to DistillBERT + Random Forest, where the latter uses 80% of the training data to train the Random Forest, while prompting ChatGPT does not use any training dataset. This demonstrates the potential of ChatGPT. Additionally, when comparing the quality of embeddings between ChatGPT and DistillBERT, the use of ChatGPT embeddings shows a significant performance increase of up to 0.25 points. Using ChatGPT Embeddings + Random Forest also outperforms simple prompting of ChatGPT.

4.1.2. Impact of prompting techniques

In this study, we explored several prompting techniques to assess their effectiveness in classifying educational-type questions, specifically zero-shot, few-shot, and chain-of-thought (CoT) approaches.

- *Zero-Shot Prompting*: This approach involves providing the language model with a task description without specific examples (Prompt-1 and Prompt-2). It leverages the model's pre-existing knowledge to classify questions based solely on the instructions given in the prompt. Our findings show that zero-shot prompting outperformed the more advanced prompting techniques across all reported statistics, including Voting, Average, and Standard Deviation.
- *Few-shot Prompting*: involves providing the model with a small number of examples alongside the task description. This technique aims to help the model better understand the task by demonstrating how examples are classified. However, our results indicated that few-shot prompting perform slightly lower than zero-shot prompting. This may be because the complexity and diversity of educational questions require a broader understanding of context than a limited set of examples can provide.
  Additionally, adding examples in the prompts increased performance variation across each prompt variation, as reflected by the higher standard deviation (Std) observed in Prompt-3 and Prompt-4. Although an F-test showed these differences were not statistically significant compared to the standard deviation of Prompt-2, the trend indicates a noticeable increase that warrants attention. Furthermore, the stability of few-shot learning can be affected by variations in example selection, order, and format, leading to inconsistent performance (Lu et al., 2022). These factors highlight the challenges of using few-shot prompting effectively for complex and varied educational content.
- *Chain-of-Thought Prompting*: enhances the model's reasoning capabilities by encouraging step-by-step problem-solving processes. In our experiments, integrating CoT with few-shot prompting did not yield significant improvements over the zero-shot approach. This is likely because the task of educational question classification does not require complex reasoning steps but rather a strong understanding of context. As a result, CoT's focus on logical reasoning does not significantly aid performance in this context.

In tasks like educational question classification, where the content is often complex and diverse, a limited set of examples in few-shot prompting may fail to capture the full context or variability required for accurate classification. As a result, the model may focus too narrowly on the specific examples provided, potentially overlooking the broader context needed for effective classification. In contrast, zero-shot prompting, which relies on the model's broader understanding of context, proves more effective. Additionally, the logical reasoning steps emphasized by Chain-of-Thought prompting are less relevant for this task, which benefits more from a comprehensive grasp of context than from complex, step-by-step reasoning.

4.1.3. Effectiveness of the voting method

The effectiveness of the Voting method for aggregating results from multiple prompts was analyzed. This method consistently outperformed not only the average of the prompt variations but also the maximum individual scores of each prompt. These findings highlight the Voting

Table 4
Comparison with Other Machine Learning Models.

No	Method	Description	F1-Score
1	Prompting ChatGPT	Zero-shot, did not use training data	0.57
2	DistillBERT Embedding + Random Forest (Faraby et al., 2022)	Use training data to train Random Forest	0.45
3	ChatGPT Embedding + Random Forest	Use training data to train Random Forest	0.7
4	Roberta + Finetuning (Faraby et al., 2022)	Use training data to finetune	0.81

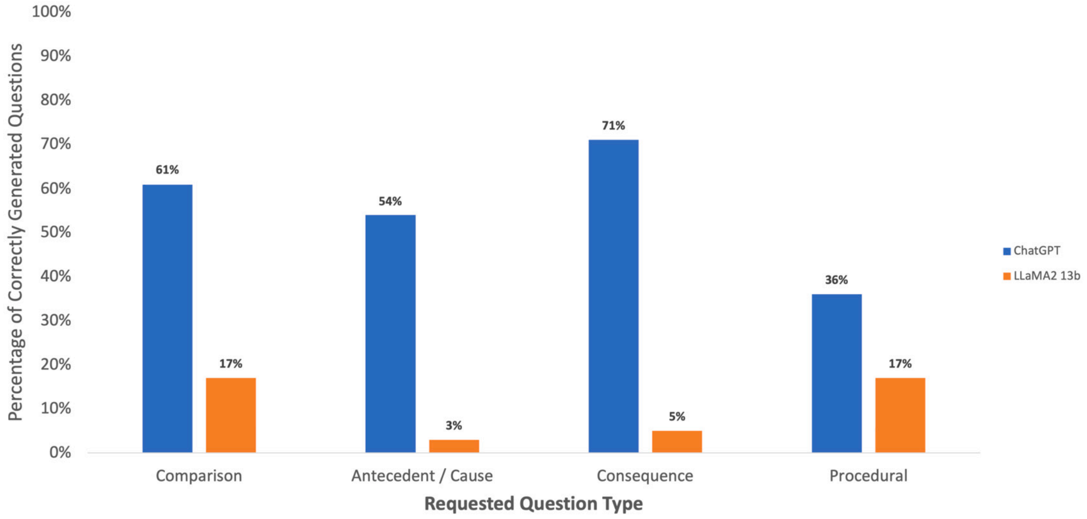


Fig. 3. Comparison of the accuracy in generating questions of the desired category by LLMs. The metric used is the percentage of questions correctly generated according to the specified type.

Table 5
Classification evaluation on self-generated questions.

Type	Against Intended Type	Against Human Labels
Comparison	0.68	0.83
Antecedent / Cause	0.58	0.79
Consequence	0.73	**0.89**
Procedural	0.35	0.54
Macro avg	0.47	0.76

Table 6
Details of the human labeling results on the questions generated by LLMs. (Comp:Comparison, Antec:Antecedent, Conseq:Consequence, Proced:Procedural, G:ChatGPT, L:LLaMA 2 13B).

Human Label	Requested Type									
	Comp		Antec		Conseq		Proced		No Type	
	G	L	G	L	G	L	G	L	G	L
Comparison	60	17	4	15	6	11	10	20	8	18
Antecedent	1	0	53	3	3	7	6	2	11	1
Consequence	0	0	0	0	70	5	0	0	2	0
Procedural	27	38	17	14	7	25	36	17	19	21
Other	11	44	24	68	13	52	48	61	59	60

method's superior ability to enhance classification accuracy and reliability.

4.1.4. *Accuracy in classifying self-generated questions*

The accuracy of classifying self-generated questions was evaluated to ensure the model's robustness, which is the ability to generalize well across different datasets and domains (Freiesleben & Grote, 2023). Specifically, we aim to ensure that the prompts created and evaluated using user-generated questions can maintain their performance when applied to questions generated by LLMs. As shown in Table 5, ChatGPT demonstrated better performance on self-generated questions with a macro-average F1-Score of 0.76, compared to 0.64 for user-generated questions. The 0.64 for user-generated questions was obtained by calculating the macro-average of the four question types from Table 3. This suggests that the model is better tuned to its own generated structures and vocabulary.

However, there were inconsistencies between the intended generation types and classification accuracy. Ideally, the classification performance should perfectly align with the types used in generating the questions, as the same model is responsible for both generation and classification. Nonetheless, inconsistencies were observed between the two. The F1-score for the intended generation types is significantly lower than that for the human labels, indicating better agreement with the human labels rather than the intended generation types. This discrepancy suggests a need for further refinement in the generation process.

4.2. *RQ2: accuracy of LLMs in generating type-specific questions*

4.2.1. *Generation accuracy performance*

In this section, we aim to evaluate the correctness of the LLMs in generating questions based on the desired types. As detailed in Subsection 3.2, the questions produced by the LLMs were labeled by two annotators into one of five predefined question types: 'comparison', 'antecedent/cause', 'consequence', 'procedural', and 'other'. The metric used here is the percentage of questions correctly generated according to the specified type. Out of a total of 1,000 samples, the inter-rater agreement, measured using Cohen's Kappa, yielded a value of 0.61, which is considered to indicate substantial agreement (McHugh, 2012).

Fig. 3 illustrates the precision of ChatGPT and LLaMA 2 13B in generating questions according to the requested type. Although ChatGPT is one of the most powerful models, with a task that does not require external knowledge other than the given input text, its performance is still relatively unsatisfactory, particularly for procedural types, where its precision is only 36%. The 'antecedent' type also only has a precision of 54%. This shows that out-of-the-box ChatGPT is relatively imprecise for this problem.

When compared to LLaMA 2 13B, ChatGPT significantly outperforms in generating questions of specific types. Although the task of question generation may seem simple, as it only involves transforming input text into questions, the specific type requests make this task challenging for smaller language models (LLMs). It can be said that LLaMA 2 13B is unable to generate questions with the specified types. It is important to note that there is a difference in the length of the input text given to LLaMA 2 13B. Due to memory limitations, the source text is limited to only 1500 tokens. This may have an impact, but considering the high precision of ChatGPT in generating questions of specific types from just a single sentence, as presented in Subsection 4.2.3, the limitation on the length of the source text should not be considered a valid reason. The detailed results of human labeling for generated questions by ChatGPT and LLaMA are provided in Table 6.

If we examine Table 6 in more detail, we can see that most of the questions generated by LLaMA are labeled as 'Other' by human annotators. Upon examination of the sample questions produced, many of them are found to be general questions that do not specifically refer to

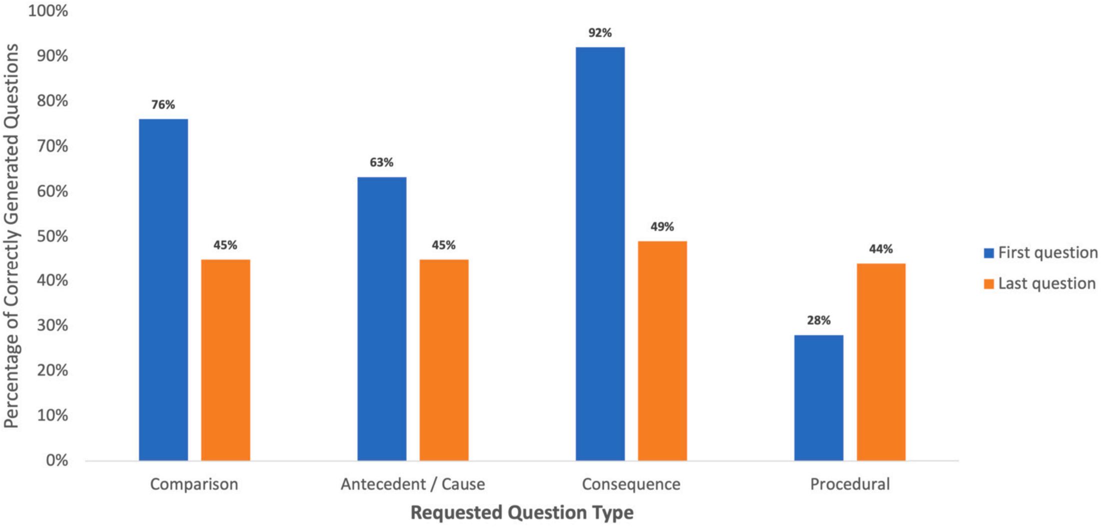


Fig. 4. Comparison of accuracy between first and last generated questions by ChatGPT across different question type.

the source text. There are also questions categorized as ‘example’ and ‘concept’ which are considered lower-order.

Here are some examples of questions generated by LLaMA2 which were supposed to be antecedent questions:

- What are the themes and organization of the textbook?
- What are some examples of successful transitions made by philosophers to other careers?
- What is the main theme of the Vedic texts?
- What is the term used to describe the esoteric doctrines about the true nature of reality beyond the realm of sense perception in the Upanishads?

Further investigation is needed to determine the extent to which finetuning can improve the capability of LLaMA 2 13B in this task.

4.2.2. *Effect of generation sequence on accuracy*

Fig. 4 compares the accuracy of the first and last generated questions across four question types: *Comparison*, *Antecedent/Cause*, *Consequence*, and *Procedural*. For *comparison* questions, accuracy drops significantly from 76% for the first questions to 45% for the last. Similarly, *antecedent/cause* questions decrease from 63% to 45%, and *consequence* questions drop sharply from 92% to 49%. Interestingly, *procedural* questions improve from 28% to 44%. These results indicate variability in the model’s performance depending on the question type and sequence position.

McNemar’s test (McNEMAR, 1947), which assesses the significance of paired data differences, was used to evaluate the accuracy differences between the first and last generated questions. Using the binomial exact method due to small sample sizes, the test revealed statistically significant differences for *comparison* ($p = 0.0004$) and *consequence* questions ($p = 4.768e\text{-}07$), suggesting accuracy varies with question position. *Antecedent* and *procedural* questions showed no significant differences ($p = 0.0636$ and $p = 0.1153$, respectively), indicating more consistent performance.

4.2.3. *Incidence of hallucinations in generated questions*

Table 7 shows the evaluation results of the generated questions and answers. The first two experiments demonstrate that LLMs are capable of generating questions with the desired type even if the type does not match the content of the source text (CT-WA: correct type - wrong answer). The condition is that LLMs are only asked to generate questions without quoting the answer. However, upon individual evaluation, all of these questions target answers that are not present in the source text. In other words, LLMs assume knowledge obtained from outside the source text.

Table 7
Assessment Results for Questions and Answers produced by LLM using Type-Inappropriate Input Text. (CT: Correct Type, CA: Correct Answer, WT: Wrong Type, WA: Wrong Answer, NComp: Not Compatible).

Prompt & Model	CT-CA	CT-WA	WT-CA	WT-WA	NComp
Comparison_GPT-3.5	0	10	0	0	0
Procedural_GPT-3.5	0	10	0	0	0
Comparison_answer_GPT-3.5	1	5	4	0	0
Procedural_answer_GPT-3.5	2	2	1	0	5
Comparison_answer_GPT-4	1	7	0	2	0
Procedural_answer_GPT-4	3	3	4	0	0

The next four experiments require LLMs to provide a snippet of the answer extracted from the source text and option “Not Compatible” as the response. This additional requirement sometimes causes LLMs to produce questions that do not match the type, but the answer pair is correct and present in the source text (WT-CA). Some also output Non-Compatible answers, which actually represents the ideal answer.

Interestingly, the requirement to quote the snippet of the answer does not always succeed, as there are still many questions and answers classified as CT-WA, where upon inspection, the provided snippet of the answer by LLMs does not actually contain the answer to the question. Furthermore, there are also CT-CA cases, which, although appearing contradictory to the definitional text, upon further examination, contain elements of procedure or comparison that are ambiguous within the definition statement.

4.3. *RQ3: quality comparison between LLM-generated and human-generated questions*

4.3.1. *Quality criteria comparison*

Table 8 presents the preferences of expert and crowdsourced evaluators between human-generated and ChatGPT-generated questions across six criteria. The criteria include clarity (C1), alignment with learning objectives (C2), stimulation of critical thinking (C3), difficulty level (C4), overall usefulness (C5), and resemblance to human-like questioning (C6).

4.3.2. *Quality comparison based on expert evaluators*

The analysis reveals that expert evaluators generally favor human-generated questions over ChatGPT-generated ones in the criteria of clarity, alignment with learning objectives, and overall usefulness. Ad-

Table 8
Preferences of Expert and Crowdsourced Evaluators Between Human-Generated and ChatGPT-Generated Questions.

Criterion	Expert Evaluator				Crowdsourced Evaluator			
	Prefer Human	Prefer Chat GPT	Confidence Interval	p-value	Prefer Human	Prefer Chat GPT	Confidence Interval	p-value
C1	0.67	0.33	[0.4171, 0.8482]	0.3018	0.46	0.54	[0.3258, 0.5971]	0.6655
C2	0.67	0.33	[0.4171, 0.8482]	0.3018	0.47	0.53	[0.3080, 0.6355]	0.8601
C3	0.33	0.67	[0.1518, 0.5829]	0.3018	0.51	0.49	[0.3675, 0.6538]	1.0000
C4	0.33	0.67	[0.1518, 0.5829]	0.3018	0.52	0.48	[0.3522, 0.6750]	1.0000
C5	0.60	0.40	[0.3575, 0.8018]	0.6072	0.44	0.56	[0.2989, 0.5896]	0.5327
C6	0.80	0.20	[0.5481, 0.9295]	**0.0352**	0.50	0.50	[0.3664, 0.6336]	1.0000

ditionally, expert evaluators tend to perceive that ChatGPT-generated questions are more difficult and better at stimulating critical thinking. However, these differences are not strong enough to be considered statistically significant across these criteria.

In contrast, the criterion of resemblance to human-like questioning stands out with a statistically significant preference for human-generated questions, as evidenced by a confidence interval that does not include 0.5 and a p-value of 0.0352. This suggests that expert evaluators significantly prefer questions generated by humans when it comes to their resemblance to human-like questioning. This finding is consistent with previous research, such as Xiao et al. (2023), which also highlighted discernible patterns distinguishing ChatGPT-generated questions from human-generated ones.

4.3.3. *Quality comparison based on crowdsourced evaluators*

On the other hand, crowdsourced evaluators did not show significant preferences for either human-generated or ChatGPT-generated questions across all criteria. Their tendency to prefer one type of question over the other was also not as strong as that of expert evaluators. These results may imply that, from the perspective of non-expert evaluators, ChatGPT-generated questions are viewed similarly to human-generated questions. This suggests that ChatGPT's performance in some areas may be comparable to human-generated content, particularly from the point of view of non-expert humans. This highlights the potential of ChatGPT in generating educational questions, demonstrating that it can produce questions perceived as comparable to those created by humans.

However, this also may be a warning that evaluations from the crowdsourced group may not be as reliable due to potential factors such as time constraints or lack of expertise. These considerations underscore the need for careful interpretation of the results, especially when comparing evaluations from expert and non-expert groups.

5. Discussion

In this section, we delve into the insights derived from our results on the effectiveness of Large Language Models (LLMs) in educational question generation and classification.

5.1. *How effectively do LLMs perform in classifying educational-type questions?*

ChatGPT demonstrated effective classification performance using zero-shot learning, relying solely on question type definitions. This approach is particularly useful in data-scarce environments where training data is limited or unavailable. The performance of ChatGPT in this context was comparable to other machine learning methods, such as distillBERT combined with Random Forest, indicating that even with minimal or no training data, zero-shot learning models can achieve high performance in classifying educational questions.

However, when ample training data is available, traditional machine learning methods that involve training or fine-tuning, such as fine-tuning models like RoBERTa, can yield better results. This suggests that while zero-shot learning offers a valuable solution for data-scarce environments, traditional machine learning approaches still hold a significant advantage in terms of accuracy and reliability. Notably, the performance of zero-shot learning in ChatGPT improved when classifying questions it generated itself, further underscoring its potential application in diverse educational settings.

5.2. *How accurate are LLMs in generating type-specific questions?*

Our evaluation of ChatGPT's accuracy in generating type-specific questions showed that it performed reasonably well, though not without challenges. While ChatGPT was able to generate questions that generally aligned with the requested types, there were instances where the generated questions did not fully match the intended types, particularly for questions generated later in a sequence. One potential solution to enhance the AQS is utilizing a question classifier as a validator or filter for automatically generated candidate questions. By addressing the initial misalignments and accuracy issues through question classification, the overall reliability and effectiveness of LLMs in AQG systems can be significantly improved.

Despite the promising capabilities of LLMs, there remains a potential for hallucination incidents, where generated questions incorporate information beyond the provided source text. In scenarios where questions are intended to stimulate exploratory thinking beyond the given text, such hallucinations might be acceptable or even desirable. However, in scenarios where the primary goal is to enhance comprehension of the reading material or to test understanding, these hallucinations can pose significant challenges. They can lead to questions that are not grounded in the provided material, thereby undermining the reliability of the LLMs for educational purposes.

To mitigate this issue, post-processing techniques are necessary. One effective approach could be to integrate a question-answering system that verifies whether the generated questions have corresponding answers within the source text. This additional layer of verification can help ensure that the questions remain relevant and accurate, thereby enhancing the overall reliability of LLM-generated content. Such measures are essential for ensuring that LLMs can be effectively utilized in educational settings without the risk of generating misleading or irrelevant questions. By addressing the hallucination problem, we can make significant strides towards the broader adoption of AI in educational content creation.

5.3. *How does the quality of questions generated by LLMs compare to those created by human experts?*

Statistical analysis revealed no significant difference in preferences between questions generated by ChatGPT and those by human experts. This finding indicates that questions generated by ChatGPT from educational source texts exhibit a quality level comparable to those generated by human experts. Evaluations by both expert groups and crowdsourced evaluators support this conclusion, suggesting that AI-generated questions can effectively mimic the quality and depth of human-generated content.

This parity in quality highlights the potential of LLMs like ChatGPT to contribute meaningfully to educational content creation, providing a scalable and efficient solution for generating educational materials. However, there is room for enhancing the robustness of this evaluation process. Increasing the number of evaluators and expanding the sample size of questions being compared could provide a more comprehensive assessment of the differences and similarities between human-generated and AI-generated questions.

Additionally, further research could explore the specific attributes that make AI-generated questions comparable to those crafted by human experts, identifying areas for improvement and refinement in AI question generation techniques. Understanding these attributes can help refine the models and improve the reliability and effectiveness of AI-generated educational content.

6. Implications for educational theories, pedagogies, and practice

Our study's findings have significant implications for educational practice, offering practical benefits for educators and institutions, particularly in the field of automatic question generation (AQG) and question classification.

1. Enhancement of Educational Content: By integrating question classification with AQG, our research supports the constructivist theory of learning, emphasizing active engagement and knowledge construction. Educators can utilize AI-generated questions to align with learning objectives, stimulating critical thinking and enhancing learner engagement. This capability allows educators to focus on facilitating deeper cognitive processing rather than solely crafting questions.
2. Personalized and Adaptive Learning: The use of zero-shot and few-shot learning techniques in our study enables the creation of personalized educational content. Educators can directly benefit from the ability to tailor questions to the specific needs and abilities of individual learners, promoting differentiated instruction and ensuring that all students are challenged appropriately.
3. Efficiency in Educational Practices: Institutions can leverage AI-generated questions to streamline the question creation process, reducing the time and effort required by educators to develop high-quality assessments. This efficiency allows educators to allocate more time to direct instructional activities and student support.
4. Monitoring and Mitigation of AI Limitations: Our findings on potential hallucinations in AI-generated questions underscore the need for careful implementation and monitoring of AI tools. Educators and institutions must ensure that generated questions are relevant and grounded in the provided material to maintain assessment integrity. By being aware of these limitations, educational practitioners can effectively integrate AI technologies into curricula and instructional strategies, maximizing their benefits while minimizing risks.

In summary, our research offers practical insights for educators and institutions, empowering them to enhance educational content, personalize learning experiences, and efficiently integrate AI technologies to improve educational outcomes.

7. Conclusion and limitation

This study has demonstrated the potential and limitations of large language models (LLMs) like ChatGPT in classifying and generating educational-type questions. Our findings show that while LLMs can effectively classify and generate questions, there are still areas requiring further refinement to match human-level quality.

The significance of this research lies not only in the technical achievements but also in its pedagogical implications. As emphasized at the beginning of this article, the ability to generate and classify questions is fundamental to the educational process. Questions drive learning by prompting critical thinking, assessing comprehension, and guiding instruction. The introduction of AI-generated questions presents an opportunity to enhance these processes by providing a scalable and adaptable solution to meet diverse educational needs.

LLMs have shown significant potential as both educational question generators and classifiers. In terms of clarity, alignment with learning objectives, difficulty level, stimulating critical thinking, and overall usefulness, LLMs with appropriate prompting can produce questions comparable to those generated by human experts. This is particularly beneficial for educational applications such as personalized learning and increasing student engagement.

However, there are areas that still require improvement. For example, when generating type-specific questions, the accuracy is not yet perfect, and there is a tendency for accuracy to decrease with the sequence of generated questions, with later questions being less accurate. Additionally, there is the issue of hallucinations, where questions may include information beyond the provided source text. This is a significant concern, especially in scenarios where questions are meant to test knowledge of specific material, highlighting the need for careful use of LLMs.

Despite these challenges, there is an opportunity to leverage LLMs as question classifiers to enhance the output of question generation systems. For instance, LLMs can serve as validators to ensure that the generated questions match the specified types. Zero-shot prompting is particularly suitable when there is little to no training data available for training a classifier. However, if ample training data is available, methods such as fine-tuning a pretrained model are more effective. Employing question classifiers after question generation is expected to make question generation systems more effective for educational use.

Future research should focus on addressing the issue of hallucinations, such as integrating a QA system to verify that the generated questions can be answered based on the source text, or ensuring that all information in the questions is present in the source text. This would make question generation systems more robust and their outputs more immediately usable without concerns over misinformation.

Lastly, we acknowledge several limitations in this study, including the relatively small number of annotators, potential bias due to annotator interpretation, the inherent ambiguity of some questions, and the limited representativeness of the dataset, which was sourced from online question-and-answer forums. Additionally, the limited variety of LLM models compared and the depth of analysis on some observed anomalies are noted. To ensure more reliable results, future research should increase the number of annotators, expand the variety of analyzed questions, and explore a broader range of LLMs to provide a more comprehensive understanding of the models' capabilities and limitations.

In light of recent advancements in LLMs, such as GPT-4 and GPT-4o, which have demonstrated superior performance in many educational tasks compared to ChatGPT 3.5, it is crucial for future studies to investigate these newer models. Evaluating their effectiveness in overcoming current limitations, particularly in reducing hallucinations and improving question accuracy, can offer significant insights.

In conclusion, this study contributes to the ongoing discussion about the role of AI in education by providing evidence that AI-generated questions, while not yet perfect, hold significant promise for enhancing pedagogical practices. By continuing to refine these technologies, we can move closer to realizing their full potential in creating engaging, effective, and personalized learning experiences.

Financial disclosure

This work was supported by Direktorat Riset, Teknologi, dan Pengabdian Kepada Masyarakat, Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi, Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia (Grant number 156/E5/PG.02.00.PT/2022) and Telkom University (Grant number 274/PNLT3/PPM/2023).

CRediT authorship contribution statement

Said Al Faraby: Writing – original draft, Project administration, Methodology, Investigation, Data curation. **Ade Romadhony:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization. **Adiwijaya:** Writing – review & editing, Supervision, Funding acquisition.

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.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used ChatGPT in order to translate and paraphrase due to being non-native speakers. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Acknowledgements

This research was financially supported by Direktorat Riset, Teknologi, dan Pengabdian Kepada Masyarakat, Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi, Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia (Grant number 156/E5/PG.02.00.PT/2022) and Telkom University (Grant number 274/PNLT3/PPM/2023).

Additionally, the Artificial Intelligence Center (AI Center) of Telkom University provided valuable support by lending computing resources, including the DGX A100.

Special thanks to Fajri Koto for his insightful suggestions, which significantly contributed to improving the robustness of the LLMs prompting in this research.

Appendix A. Appendix section

A.1. Prompts strategies

In the course of our experiments with language models (LLMs), we adhered to prompting strategies recommended by OpenAI (OpenAI, 2023), with particular focus on crafting clear instructions and incorporating reference texts. This involved several key approaches.

Clear Instruction: First, we included required details such as definition of a question type. To enhance clarity and ensure precise input demarcation, we utilized delimiters such as single quotes and triple quotes. Furthermore, we specified the sequential steps by writing prompts in blocks, which included: understanding the definition of the question, comprehending the reading material, formulating relevant questions, and outputting responses in the required format.

Provide Text Reference: In addition, we integrated reference texts into the prompts and delineated by triple quotes. Here are the implementation of the prompting techniques in one of our prompts.

```
**BEGIN PROMPT**
Comparison Question: Asking for comparison among multiple events
or concepts.

Instruction: Based on the reading section below, create some
"comparison questions" and use the following format.

Q1:
Q2:
Qn:

Here is the reading section:

"""The Most Common Types of Business Organization ...
"""
**END PROMPT**
```

Few-shot learning: To further enhance the efficacy of our prompting methods, we provided multiple examples within our question classification prompts which serves as training data for in-context *few-shot* learning and as a way to format the output. Here is the example.

```
**BEGIN PROMPT**
Classify the following question into one of
the following categories:
a. verification. Definition: Asking for the truthfulness of
   an event or a concept.
Examples:
- Is Michael Jackson an African American?
- Does a Mercedes dealer have to unlock a locked radio?

...

Question: Is it possible to be your brothers best friend
once hes married?
Category: verification

Question: When two bacteria exchange genetic information,
what is the process called?
Category: concept

...

Question: {question}
Category:
**END PROMPT**
```

Chain-of-Thought (CoT): Subsequently, we combined few-shot learning with CoT to enable LLMs to think step-by-step. Below is an example of how CoT is implemented in our prompt.

```
**BEGIN PROMPT**
....
Classify a question sentence into precisely one of the above
categories by following these steps:

1. Read the question carefully to understand its intent.
2. Compare the question's intent with the definitions
   and examples provided.
3. Determine the most appropriate category based on
   the comparison.
4. Assign the category without providing an explanation.
...
**END PROMPT**
```

By following these strategies and techniques, we ensured that our prompts were clear, detailed, and structured, resulting in more relevant and accurate responses from the LLMs.

A.2. Guideline for question type labeling

See Table A.9.

A.3. A/B evaluation form

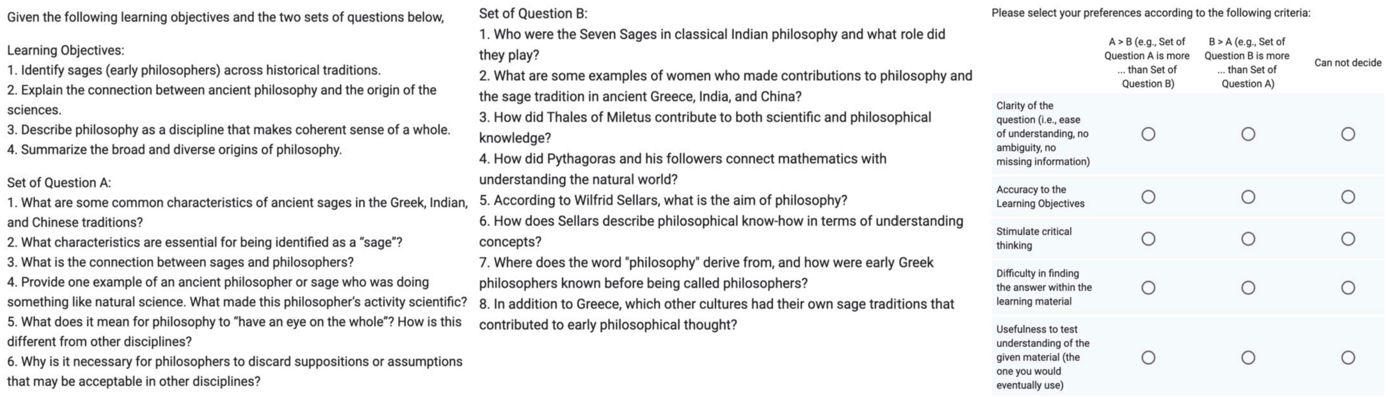
Fig. A.5 is an example of the evaluation form given to both groups of human evaluators.

A.4. List of acronyms

See Table A.10.

Table A.9
Guidelines for human annotators on question type labeling.

No	Question Types	Description	Examples
1	Comparison	Asking for comparison among multiple events or concepts.	What is the difference between sulfa (like the sulfa in antibiotics) and sulfate (like ammonium sulfate)? How does an electric violin "play" differently than an acoustic violin?
2	Antecedent	Asking for the cause or reason for an event or a concept.	How did these experiments fail? What makes nerve agents like "Novichok" so hard to produce and why can only a handful of laboratories create them? Why is the sky blue?
3	Consequence	Asking for the consequences or results of an event.	What are the negative consequences for the services if they do not evaluate their programs? What would happen if employers violate the legislation? What are the effect of climate change on the Earth's orbit/rotation?
4	Procedural	Asking for the procedures, tools, or methods by which a certain outcome is achieved.	How did the Amish resist assimilation into the current social status in the U.S? How astronomers detect a nebula when there are no stars illuminating it? How do firefighters determine the source of a fire?
5	Other	Other than 4 types above. Actually, there are 19 types of question in this scheme, but this time we want to focus only to the four types. All 19 types of question and their descriptions are listed below. When you are in doubts, you can choose this label.	



Given the following learning objectives and the two sets of questions below,

Learning Objectives:
1. Identify sages (early philosophers) across historical traditions.
2. Explain the connection between ancient philosophy and the origin of the sciences.
3. Describe philosophy as a discipline that makes coherent sense of a whole.
4. Summarize the broad and diverse origins of philosophy.

Set of Question A:
1. What are some common characteristics of ancient sages in the Greek, Indian, and Chinese traditions?
2. What characteristics are essential for being identified as a "sage"?
3. What is the connection between sages and philosophers?
4. Provide one example of an ancient philosopher or sage who was doing something like natural science. What made this philosopher's activity scientific?
5. What does it mean for philosophy to "have an eye on the whole"? How is this different from other disciplines?
6. Why is it necessary for philosophers to discard suppositions or assumptions that may be acceptable in other disciplines?

Set of Question B:
1. Who were the Seven Sages in classical Indian philosophy and what role did they play?
2. What are some examples of women who made contributions to philosophy and the sage tradition in ancient Greece, India, and China?
3. How did Thales of Miletus contribute to both scientific and philosophical knowledge?
4. How did Pythagoras and his followers connect mathematics with understanding the natural world?
5. According to Wilfrid Sellars, what is the aim of philosophy?
6. How does Sellars describe philosophical know-how in terms of understanding concepts?
7. Where does the word "philosophy" derive from, and how were early Greek philosophers known before being called philosophers?
8. In addition to Greece, which other cultures had their own sage traditions that contributed to early philosophical thought?

Please select your preferences according to the following criteria:

	A > B (e.g., Set of Question A is more ... than Set of Question B)	B > A (e.g., Set of Question B is more ... than Set of Question A)	Can not decide
Clarity of the question (i.e., ease of understanding, no ambiguity, no missing information)	○	○	○
Accuracy to the Learning Objectives	○	○	○
Stimulate critical thinking	○	○	○
Difficulty in finding the answer within the learning material	○	○	○
Usefulness to test understanding of the given material (the one you would eventually use)	○	○	○

Fig. A.5. Example of Question Quality Evaluation Form.

Table A.10
List of Acronyms and Definitions.

Acronym	Full Form	Definition
LLM(s)	Large Language Model(s)	A type of AI model designed to understand and generate human-like text.
NLP	Natural Language Processing	A field of AI focused on the interaction between computers and humans through natural language.
GPT	Generative Pre-trained Transformer	A type of LLM developed by OpenAI, known for its ability to generate coherent text.
LLaMA	Large Language Model Meta AI	An LLM designed by Meta for various NLP tasks.
PaLM	Pathways Language Model	An LLM developed by Google for improved AI language understanding.
AQG	Automatic Question Generation	The process of automatically generating questions from a given text.
CoT	Chain of Thought	A prompting technique used in LLMs to encourage step-by-step reasoning.
BLEU	Bilingual Evaluation Understudy	A metric for evaluating the quality of text by comparing with reference text.
RNN	Recurrent Neural Network	A type of neural network used for processing sequential data, including text.
POS	Part of Speech	A grammatical category of words (such as noun, verb, etc.).
SQuAD	Stanford Question Answering Dataset	A reading comprehension dataset consisting of questions posed by crowdworkers on a set of Wikipedia articles.
API	Application Programming Interface	A set of functions and protocols for building and interacting with software applications.

References

Abdelghani, R., Wang, Y. H., Yuan, X., Wang, T., et al. (2022). *GPT-3-driven pedagogical agents for training children's curious question-asking skills*. arXiv preprint arXiv.

Al Faraby, S., Adiwijaya, A., & Romadhony, A. (2023). Review on neural question generation for education purposes. *International Journal of Artificial Intelligence in Education*.

Ali, H., Chali, Y., & Hasan, S. A. (2010). Automation of question generation from sentences. In *Proceedings of QG2010: The third workshop on question generation* (pp. 58–67). academia.edu.

Bi, S., Cheng, X., Li, Y. F., Qu, L., Shen, S., Qi, G., Pan, L., & Jiang, Y. (2021). Simple or complex? Complexity-controllable question generation with soft templates and deep mixture of experts model. In *Findings of the association for computational linguistics: EMNLP 2021* (pp. 4645–4654). Punta Cana, Dominican Republic: Association for Computational Linguistics.

Bloom, B. S. (1956). *Taxonomy of educational objectives: The classification of educational goals. Cognitive domain*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (pp. 1877–1901). Curran Associates, Inc.

Cao, S., & Wang, L. (2021). Controllable open-ended question generation with a new question type ontology. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)* (pp. 6424–6439). Association for Computational Linguistics (Online).

Chan, Y. H., & Fan, Y. C. (2019). A recurrent BERT-based model for question generation. In *Proceedings of the 2nd workshop on machine reading for question answering* (pp. 154–162).

Cheng, Y., Ding, Y., Pascual, D., Richter, O., Volk, M., & Wattenhofer, R. (2021). WikiFlash: Generating flashcards from Wikipedia articles. In *AAAI 2021 workshop on AI education-35th AAAI conference on artificial intelligence (AAAI)*. tik-old.ee.ethz.ch.

Chin, C., & Osborne, J. (2008). Students' questions: A potential resource for teaching and learning science. *Studies in Science Education*, *44*, 1–39.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., … Fiedel, N. (2022). *PaLM: Scaling language modeling with pathways*.

Cooper, G. (2023). Examining science education in ChatGPT: An exploratory study of generative artificial intelligence. *Journal of Science Education and Technology*, *32*, 444–452.

Crompton, H., & Burke, D. (2024). *The educational affordances and challenges of ChatGPT: State of the field*. TechTrends.

Du, X., Shao, J., & Cardie, C. (2017). Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 1342–1352). Stroudsburg, PA, USA: Association for Computational Linguistics.

Faraby, S. A., Adiwijaya, & Romadhony, A. (2022). Educational question classification with pre-trained language models. In *2022 seventh international conference on informatics and computing (ICIC)* (pp. 1–6).

Fei, T., Heng, W. J., Toh, K. C., & Qi, T. (2003). Question classification for e-learning by artificial neural network. In *Fourth international conference on information, communications and signal processing, 2003 and the fourth Pacific Rim conference on multimedia. Proceedings of the 2003 joint, Vol. 3* (pp. 1757–1761).

Freiesleben, T., & Grote, T. (2023). Beyond generalization: A theory of robustness in machine learning. *Synthese*, *202*, 109.

Gao, Y., Bing, L., Chen, W., Lyu, M., & King, I. (2019). Difficulty controllable generation of reading comprehension questions. In *Proceedings of the twenty-eighth international joint conference on artificial intelligence, international joint conferences on artificial intelligence organization, California* (pp. 4968–4974).

Graesser, A. C., & Person, N. K. (1994). Question asking during tutoring. *American Educational Research Journal*, *31*, 104–137.

Gu, J., Zhao, H., Xu, H., Nie, L., Mei, H., & Yin, W. (2023). Robustness of learning from task instructions. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the association for computational linguistics: ACL 2023* (pp. 13935–13948). Toronto, Canada: Association for Computational Linguistics.

Haris, S. S., & Omar, N. (2012). A rule-based approach in bloom's taxonomy question classification through natural language processing. In *2012 7th international conference on computing and convergence technology (ICCCT)* (pp. 410–414).

Heilman, M., & Smith, N. A. (2010). Good question! Statistical ranking for question generation. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics* (pp. 609–617). Stroudsburg, PA, USA: Association for Computational Linguistics.

Hofstein, A., Navon, O., Kipnis, M., & Mamlok-Naaman, R. (2005). Developing students' ability to ask more and better questions resulting from inquiry-type chemistry laboratories. *Journal of Research in Science Teaching*, *42*, 791–806.

Horbach, A., Aldabe, I., Bexte, M., de Lacalle, O. L., & Maritxalar, M. (2020). Linguistic appropriateness and pedagogic usefulness of reading comprehension questions. In *Proceedings of the 12th language resources and evaluation conference* (pp. 1753–1762). aclweb.org.

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, Article 102274.

Koto, F., Aisyah, N., Li, H., & Baldwin, T. (2023). Large language models only pass primary school exams in Indonesia: A comprehensive test on IndoMMLU. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 12359–12374). Singapore: Association for Computational Linguistics.

Kumar, V., Hua, Y., Ramakrishnan, G., Qi, G., Gao, L., & Li, Y. F. (2019). Difficulty-controllable multi-hop question generation from knowledge graphs. In *The semantic Web – ISWC 2019* (pp. 382–398). Springer International Publishing.

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., & Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digit Health*, *2*, Article e0000198.

Li, Y., Lin, Z., Zhang, S., Fu, Q., Chen, B., Lou, J. G., & Chen, W. (2023). Making language models better reasoners with step-aware verifier. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 5315–5333). Toronto, Canada: Association for Computational Linguistics.

Lu, Y., Bartolo, M., Moore, A., Riedel, S., & Stenetorp, P. (2022). Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 8086–8098). Dublin, Ireland: Association for Computational Linguistics.

Marzi, G., Balzano, M., & Marchiori, D. (2024). K-alpha calculator–Krippendorff's alpha calculator: A user-friendly tool for computing Krippendorff's alpha inter-rater reliability coefficient. *MethodsX*, *12*, Article 102545.

Mathur, N., Baldwin, T., & Cohn, T. (2020). Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 4984–4997). Association for Computational Linguistics (Online).

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*, 276–282.

McNEMAR, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, *12*, 153–157.

Mitkov, R., & Le An, H. (2003). Computer-aided generation of multiple-choice tests. In *Proceedings of the HLT-NAACL 03 workshop on building educational applications using natural language processing* (pp. 17–22).

Mizrahi, M., Kaplan, G., Malkin, D., Dror, R., Shahaf, D., & Stanovsky, G. (2023). State of what art? A call for multi-prompt LLM evaluation. Retrieved from arXiv:2401.00595.

Mohammed, M., & Omar, N. (2020). Question classification based on bloom's taxonomy cognitive domain using modified TF-IDF and word2vec. *PLoS ONE*, *15*, Article e0230442.

Mostow, J., & Chen, W. (2009). Generating instruction automatically for the reading strategy of self-questioning. In *Proceedings of the 2009 conference on artificial intelligence in education: Building learning systems that care: From knowledge representation to affective modelling* (pp. 465–472). Amsterdam, the Netherlands, the Netherlands: IOS Press.

Nappi, J. S. (2017). The importance of questioning in developing critical thinking skills. *Delta Kappa Gamma Bulletin*, *84*, 30.

Nov, O., Singh, N., & Mann, D. (2023). Putting ChatGPT's medical advice to the (Turing) test: Survey study. *JMIR Medical Education*, *9*, Article e46939.

Olney, A. M. (2023). Generating multiple choice questions from a textbook: Llms match human performance on most metrics. In *AIED workshops*.

OpenAI (2023). *Prompt engineering*. OpenAI.

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 2383–2392).

Sekulić, I., Aliannejadi, M., & Crestani, F. (2021). Towards facet-driven generation of clarifying questions for conversational search. In *Proceedings of the 2021 ACM SIGIR international conference on theory of information retrieval* (pp. 167–175). New York, NY, USA: Association for Computing Machinery.

Srivastava, M., & Goodman, N. (2021). Question generation for adaptive education. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 2: Short papers)* (pp. 692–701). Association for Computational Linguistics (Online).

Sultan, M. A., Chandel, S., Astudillo, R. F., & Castelli, V. (2020). On the importance of diversity in question generation for QA. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 5651–5656).

Sun, X., Li, X., Li, J., Wu, F., Guo, S., Zhang, T., & Wang, G. (2023). Text classification via large language models. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the association for computational linguistics: EMNLP 2023* (pp. 8990–9005). Singapore: Association for Computational Linguistics.

Syed, R., Collins-Thompson, K., Bennett, P. N., Teng, M., Williams, S., Tay, D. W. W., & Iqbal, S. (2020). Improving learning outcomes with gaze tracking and automatic question generation. In *Proceedings of the Web conference 2020* (pp. 1693–1703). New York, NY, USA: Association for Computing Machinery.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., … Scialom, T. (2023). *Llama 2: Open foundation and fine-tuned chat models*.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2022). Self-consistency improves chain of thought reasoning in language models. Retrieved from arXiv:2203.11171.

Xiao, C., Xu, S. X., Zhang, K., Wang, Y., & Xia, L. (2023). Evaluating reading comprehension exercises generated by LLMs: A showcase of ChatGPT in education applications. In E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, & T. Zesch (Eds.), *Proceedings of the 18th workshop on innovative use of NLP for building educational applications (BEA 2023)* (pp. 610–625). Toronto, Canada: Association for Computational Linguistics.

Xiao, D., Zhang, H., Li, Y., Sun, Y., Tian, H., Wu, H., & Wang, H. (2020). ERNIE-GEN: An enhanced multi-flow pre-training and fine-tuning framework for natural language generation. In *Proceedings of the twenty-ninth international joint conference on artificial intelligence, international joint conferences on artificial intelligence organization, California*.

Yahya, A. A., & Osman, A. (2011). Automatic classification of questions into bloom's cognitive levels using support vector machines. In *The international arab conference on information technology, 212.26.74.23* (pp. 1–6).

Zhai, X. (2023). ChatGPT for next generation science learning. *XRDS*, *29*, 42–46.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., … Wen, J. R. (2023). A survey of large language models. Retrieved from arXiv:2303.18223v11.

Zulqarnain, M., Khalaf Zager Alsaedi, A., Ghazali, R., Ghouse, M. G., Sharif, W., & Aida Husaini, N. (2021). A comparative analysis on question classification task based on deep learning approaches. *PeerJ Computer Science*, *7*, Article e570.