Computers and Education: Artificial Intelligence 9 (2025) 100440

Contents lists available at ScienceDirect

Computers and Education: Artificial Intelligence

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

Comparative analysis of NLP-driven MCQ generators from text sources

Asmae Azzi [a], Ferenc Erdős [a,*], Richárd Németh [a], Vijayakumar Varadarajan [b,e], Stephen Afrifa [c,d]

[a] Széchenyi István University, Győr, Hungary
[b] University of Technology Sydney, Australia
[c] School of Electrical and Information Engineering, Tianjin University, Tianjin, China
[d] Department of Information Technology and Decision Sciences, University of Energy and Natural Resources, Sunyani, Ghana
[e] Swiss School of Business and Management, Geneva, Switzerland

ARTICLE INFO

Keywords:
Natural language processing
Multiple-choice questions
Artificial intelligence
Large language models
MCQ generators
LLM in education

ABSTRACT

The application of learning sciences with technology has been shown to boost learner interactions, yet the potential of advanced tool, particularly those that leverage Natural Language Processing (NLP), still very much untapped in learning contexts. This paper speaks to this age-old problem of generating quality Multiple-Choice questions (MCQs) – a prevalent but time-consuming mode of assessment – via the suggested comprehensive comparison study of template-based AI solutions. The study contrasts general-purpose Large Language Models (LLMs) with specialized MCQ-focused AI programs. The scientific approach employed was quite stringent, where each of the software applications was benchmarked using a common dataset of text across varying levels of complexity and topic. Results indicate that general-purpose LLMs, especially DeepSeek and ChatGPT, consistently present higher performance and reliability, especially when processing complex textual content. Whereas specialized tools offer distinctive formatting options, they exhibit decreasing performance as texts become more complex and signify strong operation constriction at the free versions. Developing solid and effective distractors turned out to be a complicated task for all the tested tools. We conclude the paper by presenting a standardized assessment model, making evidence-based recommendations for developers and teachers, and suggesting ways to incorporate various AI capabilities into modern educational assessment effectively.

1. Introduction

MCQs are examination questions in which the person responding is presented with different possible answers to a question from which they must select the correct answer (Haladyna & Downing, 2004). Unlike dichotomous (binary type) questions, where response options are limited to 'yes/no' or 'true/false' (Sha, 2022), these multiple-choice questions involve knowledge rather than luck in choosing the correct answer, as the respondents must choose between several alternatives. MCQs are among the most prevalent forms of quizzes used in education (Collins, 2006). However, it might be time-consuming to generate high-quality MCQs, especially for educators with a heavy workload.

While multiple studies have examined MCQ generation tools, the appearance of new LLMs (Large Language Models) like ChatGPT, Gemini, and DeepSeek as well as specialized MCQ generation solutions, necessitates new evaluations. Girish and Rekha Prabhu (Prabhu, G. & Prabhu, R., 2023), and Bharatha et al. (Bhratha et al., 2024) evaluated the MCQ generation capabilities of ChatGPT using statistical indicators (mean, standard deviation, and Bloom's Taxonomy). In addition, there have also been studies comparing the performance of ChatGPT, Bard (now Gemini), and Bing using mean and Cohen's Kappa coefficient (Agarwal et al., 2023), and assessing the knowledge of ChatGPT and Gemini using revised Bloom's Taxonomy (Sallam et al., 2024). LLMs have been successfully used for question generation by Kopi Aliena's Lab (Firdaus et al., 2024) and Volodymyr Mavrych's Lab (Bolgova et al., 2023). However, we found that there is no truly comprehensive study in the literature on benchmarking MCQ generators with performance comparison.

When generating questions from plain texts, the tools usually transform the declarative sentences in the text into questions by focusing on the main ideas of each sentence based on selected keywords. For instance, "AUTOQUEST" was among the first tools that used a similar technique. Although it is no longer available, it still represents an important example of MCQ generation programs. It was mainly used for

* Corresponding author. Széchenyi István University, Egyetem tér 1. Győr, Hungary.
E-mail address: erdosf@sze.hu (F. Erdős).

https://doi.org/10.1016/j.caeai.2025.100440
Received 14 March 2025; Received in revised form 25 May 2025; Accepted 13 June 2025
Available online 21 June 2025

testing students' understanding after reading a text by generating a set of questions for each sentence. The program's vocabulary was based on a dictionary of different grammatical categories ranging from pronouns to verb stems (Wolfe, 1976). The first step "AUTOQUEST" follows is to preprocess the given sentence, then develop a set of patterns based on the sentence's characteristics, and finally post-process those questions before outputting them.

It is still challenging for educators and quiz creators to select the best tool, as current studies lack a comprehensive benchmark that includes recent models. An updated and standardized analysis that evaluates various MCQ-generating tools across key performance indicators is urgently needed, given the growing reliance on AI in business and educational settings.

1.1. Organization of the paper

After a brief introduction, our paper reviews the most important technologies and practical applications involved in the research, such as the concept of MCQ, generative Large Language Models (ChatGP, Gemini, and DeepSeek), and question generation programs. After some historical overview and an introduction to the milestones of AI models and increasingly advanced neural networks, we move on to the presentation of the research methodology, in which the data set we have compiled, the survey details, and the method of participant selection are described. The obtained evaluations have been assessed using the comparative analysis methodology we have devised, using the appropriate indicators.

In the study, great emphasis has also been placed on the visualization of the scores obtained, for ease of comparison; the latter being the main part of the paper. The answers to the questions asked are also assessed in terms of quality and reliability. Following the practical values, we also subject each target system to secondary evaluation (such as user-friendliness or ease of use), as well as the limitations of each tool are described.

Before the conclusion, we provide insight into the challenges associated with the tools and formulate suggestions for making MCQ generators more effective.

The study concludes with the results of the primary and secondary evaluations established during the research, identifying further development directions.

1.2. Significance of the study and the aim of the research

In particular, our paper seeks to fill a knowledge gap in the existing literature by comparing AI-driven tools, presenting a uniform benchmark for a full comparison of a wide range of models. The study has three main scientific contributions: 1) It provides a comparative analysis of different multiple-choice question generators and the representatives of Large Language Models. These are compared across a wide range of educational topics and evaluated based on properties such as question quality, correctness of answers, and appropriateness of confounding factors.; 2) Provides a standardised evaluation framework that can be applied in future AI-driven MCQs and LLM models. The latter will allow for continuous comparative analysis as new models emerge. Finally, 3) the research highlights the main challenges of current AI-driven software in the development of multiple-choice question generators, including language adaptability, content adequacy, and the ability to generate high-quality distractors.

With this research, our goal is to perform a comparative analysis of current AI-driven tools for automatic MCQ generation, evaluating their capabilities, limitations, and performance to assist quiz creators and instructors in selecting the best option. We aim to suggest evidence-based recommendations to help quiz creators select the best tools that meet their expectations and specific needs.

Although previous research (Kurdi et al., 2020) classified MCQ generation tools, our paper provides a standardized benchmark to evaluate these tools comprehensively by presenting a structured evaluation of multiple models.

Building on this, the following research questions guide our study.

RQ1: How do the existing AI MCQ generators compare in terms of question quality, answer accuracy, and distractors' relevance?
RQ2: How can MCQ-generating tools be evaluated using a standardized evaluation framework?
RQ3: Compared to domain-specific MCQ generators, what are the strengths and weaknesses of general-purpose LLMs?
RQ4: What are the main challenges in AI-based MCQ generation, and what suggestions are there for quiz creators?

1.3. Background information

In addition to the usage a largely uniform approach to MCQ generation, most tools available on today's market share the same base architecture that was first presented by Vaswani et al., in 2017, which is the so-called transformer architecture. Recently, RNNs (Recurrent Neural Networks), Long Short-Term Memory Models (LSTM), and Gated Recurrent Units (GRU) were used for word processing tasks. However, with the introduction of the attention mechanism and positional encoding for faster than sequential processing, transformer models have proven to be more effective and faster (Lu et al., 2020). Self-attention, in this context, can be defined as an attention system that links different points in some sequence to compute its illustration (Vaswani et al., 2017). This solution allows the model to dynamically weigh the different parts of the input sequence to better understand the contextual information. Compared to other neural network architectures that can be used for similar purposes (LSTM and RNN), transformers have the big advantage that they can be trained and operated on sequential data faster (Ye & Pandarinath, 2021), as they do not require traditional sequential processing, which greatly reduces the learning time, making the model more efficient to use. However, even though most of today's tools that treat NLP tasks are based on the transformer architecture, we can notice some differences. We can distinguish three types of models: encoder-only, decoder-only, and encoder-decoder model. For instance, BERT (Bidirectional Encoder Representations from Transformers) is based on an encoder-only architecture, it is mainly a stack of transformer encoder layers that include many self-attention heads (Rogers et al., 2020). Rogers et al. (2020) explain the workflow in BERT as each head computes key, value, and query vectors (Rogers et al., 2020) for every input token present in some pattern, which is utilized to produce a weighted depiction. Every output coming from the same layer's head is merged and routed via a fully linked layer (Rogers et al., 2020). Every layer is preceded by layer normalization and completed with a skip connection. Programs that are based on encoder-only architecture like BERT are usually used for tasks that focus on sequence embedding, text classification, information retrieval, or anomaly detection. Additionally, they can also be used for MCQ generation, especially if integrated with other algorithms. For the second type, decoder-only architecture programs are based on the transformer architecture but only use the decoder part. These programs consist of a stack of transformer decoder layers that include: "multi-head self-attention" and "position-wise feed forward networks" (Vaswani et al., 2017). For example, GPT-2 follows the same architecture and can be used to perform a multitude of NLP tasks ranging from translation to question answering. Furthermore, GPT-3 and later models also follow the same architecture but can output higher-quality results mainly because they have larger parameters and were trained and fine-tuned on more diverse datasets (Floridi & Chiriatti, 2020). For instance, GPT-3 used 175 billion parameters while GPT-2 used only 1.5 billion of them (Floridi & Chiriatti, 2020). Finally, it is worth mentioning the tools that use both parts of the transformer architecture: the encoder and decoder such as T5 (Text-To-Text Transfer Transformer) which treats every text-processing task as a "text to text" situation (Raffel et al., 2020). The T5 model can perform a variety of

tasks such as translation, summarization, grammar correction, and semantic similarity detection. Additionally, the pre-training of the T5 model was performed on large datasets such as: "WMT' 14 English to French" (Raffel et al., 2020) and "Colossal Crawled Corpus" (Raffel et al., 2020), while the fine-tuning was done on diverse NLP tasks like translation and summarization (Raffel et al., 2020). The way those two aspects were completed differentiates the T5 model from the original transformer architecture. The T5 model has tremendous promise for a range of NLP applications, including MCQ production (Raffel et al., 2020). The flexibility of this tool allowed developers and researchers to create software solutions that make use of this model.

This correlates with the studies conducted by Kurdi, Haffari, and Rasooli (Kurdi et al., 2020) who provided a thorough review of the current state of automatic MCQ generation from an input text. They divided the MCQ-generating techniques into two groups: rule-based methods and data-driven methods (Kurdi et al., 2020). For the first group, the systems employ a set of manually produced rule-based guidelines, whereas the second type uses machine learning. The authors explore a variety of automatic MCQ-generating techniques, including open-source programs, research datasets, and commercial solutions such as Quillionz, Quizizz, Quizlet, and others. Additionally, the authors identified major challenges in the field, such as producing high-quality MCQs across various domains and languages (Kurdi et al., 2020).

Beyond their technical design, multiple-choice questions continue to hold considerable pedagogical value in modern education. Meta-analytic evidence indicates that low-stakes MCQ quizzes, particularly when they are accompanied by timely feedback, contribute significantly to long-term knowledge retention and reduce test-related anxiety (Yang et al., 2023). This is a phenomenon widely referred to as the testing effect (Yang et al., 2023). This dual role, as both an instrument for assessment of learning and assessment for learning, makes MCQs especially suitable for formative assessment contexts, where iterative feedback is essential to the learning process.

The alignment between MCQ design and established educational theories reinforces their value. In a constructivist framework and under Bloom's revised taxonomy, well-crafted MCQs promote active engagement and support higher-order thinking. Haataja et al. (2023) show that such items can effectively target cognitive levels like analyze, evaluate, and create. Likewise, Cognitive Load Theory (CLT) emphasises the importance of clarity and brevity in question design to reduce unnecessary cognitive strain, particularly for beginner learners (Barbieri & Rodrigues, 2025).

MCQs also align with retrieval-based learning. Yang et al. (2023) found that repeated MCQ practice, especially when combined with feedback, improves long-term retention and learning transfer. This makes MCQ quizzes effective tools in spaced practice.

Recent evidence highlights the pedagogical promise of AI-generated MCQs. Bitew et al. (2024) report that context-aware distractor generation produced high-quality items in over 30 % of cases, outperforming template-based approaches. Newton and Xiromeriti (2024) show that GPT-4 can pass high-stakes MCQ exams (unlike GPT-3.5), which raises integrity concerns. Noda et al. (2025) further demonstrate that GPT-4 enhanced physicians' performance on domain-specific MCQs by nearly 20 %, indicating its potential in expert-level assessment.

Even though the previously published papers compared some tools that automatically generate MCQs, novel large language model-based generative AI technologies like ChatGPT or Google Gemini, and other MCQ-specific solutions introduce a new and important dimension that remains underexplored in the current literature.

1.4. Scope and limitations

Three trends dominate modern MCQ research. First, large-language-model (LLM) pipelines now automate the entire item-authoring workflow, generating stems, correct responses, and distractors that span all six Bloom categories with near-expert fluency; ChatGPT, for example, answered psychosomatic-medicine questions across the taxonomy as accurately as physicians in a mixed-methods study (Herrmann-Werner et al., 2024). Second, context-aware distractor reuse – in which high-quality options from existing items are retrieved and adapted – has raised teacher-rated item quality by 30 % compared with static templates (Bitew et al., 2024). Third, computerised-adaptive testing engines increasingly pair transformer-generated items with on-the-fly difficulty calibration, halving test length while preserving reliability (Ebenbeck et al., 2024).

These advances arrive with new risks. GPT-4 already meets or exceeds pass marks on many high-stakes examinations, raising questions about validity and security (Newton & Xiromeriti, 2024). Item-response simulations also show that model-generated response patterns can differ from those of diverse student groups, calling for explicit fairness checks during calibration (Liu et al., 2025). Moreover, even the best distractor-reuse pipeline still produces "expert-ready" items in only one case out of three, keeping human vetting squarely in the loop (Bitew et al., 2024).

In the future, a growing body of work on explainable-AI dashboards aims to flag ambiguous wording and surface rationales for each correct answer, thereby supporting both teachers and learners (Khosravi et al., 2022). Psychometric studies using LLM "proxy respondents" suggest a feasible route to faster, low-cost item calibration (Bhandari et al., 2024). While these innovations lie beyond the empirical scope of the present comparison, they set important directions for future MCQ-based assessment.

2. Presentation of tools

In this comparative analysis, we have chosen a range of tools that can perform the MCQ generation task. These tools were divided into two groups: large language model-based chatbots and MCQ-specific AI solutions. For the first group, we chose ChatGPT, Gemini, and DeepSeek, while for the second group, we chose QuizGecko and Questgen. These tools are notable examples of current AI-driven MCQ generators since they rely on sophisticated language models. These programs are designed to generate MCQs from textual information and they represent specific solutions in the subject.

The first tool, ChatGPT is a chatbot that has gained enormous popularity; it has a free plan and different payable plans. From the middle of 2024, a faster and more accurate GPT-4o model became available for free with message limits. This tool was built on a decoder-only architecture and was trained to respond to a prompt by following an instruction and providing a thorough answer. ChatGPT has been trained using RLHF (Reinforcement Learning From Human Feedback), which uses similar approaches to InstructGPT which is a variant of GPT-3. ChatGPT is a valuable tool for researchers at all phases of a study, providing pertinent information, direction, and assistance to maximize productivity (Azaria et al., 2023). The tool can revolutionize the education field by creating personalized learning experiences and automating diverse tasks performed by educators (Azaria et al., 2023). However, like similar gen AI tools, ChatGPT has its limitations that may result in incorrect or biased answers which can negatively impact the student's learning. Furthermore, ChatGPT may not comprehend the nuances of human language, which might lead to misunderstandings (Azaria et al., 2023).

The next tool, Google Gemini (formerly Bard) was initially built on top of Google AI's LaMDA (Language Models for Dialog Applications), which is a set of transformer-based neural language models known for their scalability, with up to 137 billion parameters (Ahmed et al., 2023). In 2022, Google conducted a study that proved LaMDA's capabilities in enhancing conversational AI programs. It produced coherent and context-aware responses using techniques like response creation, safety filtering, and others (Ahmed et al., 2023). Gemini supports a variety of language tasks, including text generation, translation, creative writing,

and answering user queries (Saeidnia, 2023). It uses a decoder-only transformer architecture with some modifications to the original design (Chowdhery et al., 2022). The free version at the time of the research (Gemini 2.0 Flash, Q1/2025) is capable of accessing real-time data, papers, and websites, which extends its range of data retrieval and interactive tasks.

The third AI model we used is DeepSeek. Its first model (Coder), released in November 2023, initially used the same decoder-based models built on the Transformer architecture (Zhu et al., 2024), with a slightly different approach. Among several other changes, it used SwiGLU as the activation function for the Feed-Forward Network, and the so-called Grouped Query Attention solution instead of the Multi-Head Attention approach (Bi et al., 2024). Its disruptive model, DeepSeek-R1, which was launched in January 2025, now uses a new architecture, Mixture-of-Experts, which can optimize computational efficiency without compromising performance. The flexibility mentioned earlier has become even more pronounced, with most of the language models being well-suited for generating questions and answers. What makes it particularly attractive is that it is much more cost-effective (Gaur, 2025), with all its features for free in Q1 2025.

QuizGecko is an AI question generator that generates different types of questions ranging from MCQs to fill-in-the-blank style questions (QuizGecko, 2023). It offers a free and paid version of the system and the ability to customize, analyze, and publish the generated quizzes. Additionally, this tool allows users to generate quizzes from YouTube videos too.

Finally, the last tool, QuestGen is an AI-powered quiz generator that can create different types of quizzes such as MCQs, true or false questions, fill-in-the-blank, or higher-order questions (Questgen, 2023). The quizzes can also be customized based on users' needs whether in terms of difficulty, number of questions generated, or the type of quizzes. This tool also offers a free and a paid version depending on the needs of the user. Questgen employs three T5 models to keep questions relevant. One for generating Boolean questions, one for MCQs, and paraphrasing, and one for answering questions (gouthamg, 2023).

3. Research methodology

3.1. Approach

In our research, the dataset collected was subjected to an exploratory data analysis (EDA) to identify trends, patterns, and correlations in the data. This type of analysis aims to examine the data in terms of distribution, outliers, and anomalies, and then make the results easily interpretable for the audience through data visualization (Komorowski et al., 2016). EDA uses several visual techniques to graphically display these statistics, such as histograms, box plots, scatter plots, line plots, or heat maps, to summarize the main characteristics of a dataset (Deming et al., 2018).

The five tools tested in this study were grouped into two categories: (1) general-purpose large language models: ChatGPT (GPT-4o model), Gemini (2.0 Flash model), and DeepSeek (R1 model), and (2) MCQ-specific tools: QuizGecko and Questgen. While both sets were evaluated on the same input data, distinctions between these categories were considered when interpreting the results.

We intentionally chose to use plain input texts, rather than pre-existing MCQ datasets, to fairly evaluate each tool's ability to independently generate multiple-choice questions. Using pre-authored MCQs could introduce bias, as these often include structural or pedagogical cues not present in typical educational texts. Our approach ensured a standardized and unbiased benchmark across tools.

The performance of five different tools (QuizGecko, QuestGen, Gemini, ChatGPT, and DeepSeek) was investigated in three different categories: 1) Quality of the Questions; 2) Quality of the Right Answer; 3) Quality of the Distractors. We looked at which quiz generator or LLM model generated the best questions, gave the most accurate answers on average, and produced the highest quality distractors. We also identified which tool produced the most accurate results, which texts had the most accurate answers on average, and which texts had the largest discrepancies. To quantify these, Python code was written to calculate the corresponding means and variances using various Python modules such as Numpy, Seaborn, and Pandas.

3.2. Dataset

Recognizing the importance of the dataset's role in this research, we started to meticulously search for the relevant data. To address RQ2, we decided to create our own dataset to guarantee unbiased analysis and establish a standardized benchmark for evaluating MCQ generation tools. This dataset contains four different texts that can be used to test the tools, current or future. The texts are of different CEFR (Common European Framework of Reference for Languages) levels, ranging from the beginner stage A2 to the advanced one C2. We have chosen the articles from two main sources: the British Council website and Wikipedia. The first two texts from the British Council website were pre-classified at the A2 and B2 CEFR levels. However, for the Wikipedia texts (Text 3 and Text 4), we developed a multi-step evaluation process to assign appropriate CEFR levels. We applied three readability metrics: Flesch Reading Ease, Flesch-Kincaid Grade Level, and Dale-Chall Score using the Textstat Python library. The resulting scores shown in Table 1, were used to align the texts with the C1 and C2 levels, respectively.

The inclusion of materials from these two sources highlights the significance of combining academically oriented texts with more broad, publicly sourced material. The combination of these two provides a complete depiction of various styles and complexities, which enhances the dataset considerably. As the primary goal of our research is to emphasize adaptability, the most effective tools should generate MCQs from both academic (e.g., British Council) and general sources (e.g., Wikipedia). Additionally, we purposely chose different subject areas, such as education, zoology, computer science, and astronomy, to further enhance our dataset and guarantee a high-quality evaluation of each tool.

The dataset development takes different learning environments into particular consideration. The chosen texts' complexity from A2 to C2 level reflects actual educational settings where teachers evaluate students across different grade levels and language proficiencies. The diversity in complexity and topics discussed in the dataset's texts allows for a robust examination of the tools' capabilities in generating MCQs depending on the level of difficulty. By using this standardized dataset, researchers and practitioners can evaluate not only the ability of current and future tools to generate MCQs but also their capability to recognize and adapt to each text's difficulty.

Furthermore, using our dataset offers two major benefits. First, it establishes a uniform and consistent standard for measuring the effectiveness of any MCQ-generating tool, enabling insightful comparisons between various tools and periods. Second, as the dataset is aligned with CEFR, the results of the assessment using this benchmark have broader relevance in the global educational system.

3.3. Methodology

Before starting the comparative analysis of the available tools, it is important to define clear evaluation metrics to measure the efficiency of each application. These metrics serve as a standardized evaluation

Table 1
Readability metrics for Text 3 and Text 4.

Text Number	Flesch Reading Ease	Flesch-Kincaid Grade Level	Dale-Chall Score
Text 3	52.49	10.29	11.67
Text 4	28.32	14.61	12.63

framework that can be universally applied to any MCQ generator, ensuring that the assessment is objective and reproducible. The different parts of the MCQ are the basis of the evaluation.

The assessment of the quality of the questions generated represents the first primary metric in our evaluation. This examination checks the clarity, relevance, and difficulty level of the questions generated by each tool. To have an efficient MCQ quiz, we should have cohesive and clear questions that match the complexity of the text. In this way, participants can be efficiently assessed on their learning depending on their English level and the topic covered. This metric allows us to differentiate between the tools that can only generate basic questions and the more sophisticated ones.

The quality of the correct solution provided by the tools is also an important metric as it often represents the keyword extracted from the source sentence and encapsulates its central idea. A good-quality keyword is one that is appropriate and relevant to the context which allows a solid basis for the MCQ. Having a good quality answer is crucial in MCQ quizzes as they represent the basis of the participants' assessment; for this reason, it is important to avoid any false, misleading, or confusing answers.

Furthermore, the quality of the distractors generated is another essential evaluation metric. The identification of distractors is the cornerstone of designing MCQs, as these must be both plausible and yet incorrect (Kumar et al., 2023). The role of distractors is to raise the difficulty level for participants taking the test to assess their knowledge and spot their learning deficiencies. For this reason, it is important to have reasonable yet inaccurate distractors that force learners to distinguish between small discrepancies. In the case of weak distractors, it is possible to mistakenly reveal the right answer which defeats the MCQ's goal. So, the tool's ability to produce contextually relevant, difficult, and varied distractors demonstrates its expertise and value.

To ensure the reliability of the expert-based evaluation, we examined the standard deviation of the individual scores assigned by the ten evaluators for each of the three primary metrics: question quality, correct answers, and distractors. Across all tools and texts, the standard deviations were generally low, typically ranging from 0.3 to 0.6 on a 5-point scale. This indicates a good level of agreement among the evaluators and reinforces the objectivity of the assessment process.

We should also remember that performance alone can be misleading if the tool under examination is not capable of consistent performance. A model that can perform at an average level on the majority of texts and excel on a few texts, but cannot consistently generate a good solution, can hardly be expected to meet our needs.

Moreover, there are some secondary evaluation metrics that we think are important to include in our evaluation as they influence the user experience and might impact positively or negatively their overall opinion about every tool. First, the ease-of-use metric allows us to evaluate the practicality of the program by assessing its user-friendliness. This is very important for different users of these tools as they can be from different backgrounds concerning AI but would like to automatically generate MCQs. When users are faced with a software solution that is easy to use, their experience while using the application becomes more enjoyable and fruitful (Lewis, 2014). For this reason, it is important to include the ease-of-use metric in the evaluation. Second, as time is a critical factor in any sector, it is crucial to include the speed of generation of questions for each tool. The goal of these programs is to help educators in the creation of MCQs which includes reducing the time needed for content development especially when handling a significant number of inquiries.

Finally, some tools present some limitations which can create problems in use. For instance, as test takers can vary in English proficiency, it is important to keep the versatility and adaptability when taking an input text. This includes reducing the requirements regarding words or characters. For this reason, we included the word or character limitation in our evaluation.

The assessment metrics mentioned before, along with our carefully selected dataset, create a uniform and standardized framework that readers and future researchers may use to evaluate any MCQ generation tool.

3.4. Collaborative evaluation

After defining the evaluation metrics, we proceeded with a collaborative evaluation to ensure the robustness of our assessment and address Research Question 2. Originally, a sample of 15 potential participants was examined. A stringent screening process, involving the exclusion of those who had come in with negligence in their initial assessments, led to a final selection of 10 highly qualified experts. This rigorous selection process ensured that only the most competent and trustworthy judgments found their way into the study findings, thus strengthening their individual credibility, considering that they were quite limited in number. We ensured that they all had at least 10 years of educational experience (on average, 19.6 years spent in a university environment) and had experience with quiz generators, i.e., using MCQ tools in their teaching activities. The mean age of the respondents was 44.3 years, partly because one of the participants was a professor emeritus in his 70s. Except for an assistant lecturer (PhD candidate), all of them had a doctoral degree. The 10 respondents came from a total of 3 universities, all of them from the field of technology. Men were slightly overrepresented: six men and four women participated in the evaluation. Every participant was asked to evaluate the tools based on the primary metrics, which included: the quality of every question, the quality of each set of distractors, and the quality of the right answer. Participants were not told which questions, right answers, or distractors were made by which tool, as we did not want their opinions to be influenced by knowing this. The assessment employed a 0 to 5 rating system, as shown in Table 2, which represents the details of the scoring system used for the evaluation metrics previously mentioned.

This collaborative approach is a unique feature of this work. The primary metrics and rating techniques described in Table 2 provide a standardized format that others can employ to assess any MCQ generator, thereby ensuring reproducibility in future studies and enabling consistent evaluation for varied systems.

We used a systematic data collection technique to improve the collaborative evaluation approach. The participants were sent an Excel sheet meticulously built up for this task. This document had four texts paired with the MCQs generated by each tool. The form contained the primary evaluation metrics organized to facilitate efficient assessment, with a space for every tool-text combination. This helped the participants to go through the evaluation easily, rate the quality of each MCQ in a structured manner, and record their remarks if needed.

We generated the MCQs for the four texts using each software system. During this testing phase, we analyzed each tool's user-friendliness and limitations. Additionally, this individual testing allowed us to present the participants with identical sets of MCQs to ensure an equitable evaluation.

The collaborative evaluation not only gave a comprehensive picture of the tools' capacities but also added to the reliability. The involvement of experts in the field of education guaranteed that the evaluation was accurate and that its outcomes were founded on solid knowledge.

4. Comparative analysis

In this section, we address RQ1 by conducting a detailed comparative analysis of the different tools that can automatically generate MCQs, all of which are powered by NLP, to evaluate their effectiveness in this area. The comparison examines available tools on the market such as ChatGPT, Google Gemini, DeepSeek, Questgen, and QuizGecko. Using this comprehensive evaluation, we intend to present a complete overview of the tools' abilities, strengths, and weaknesses, providing insight into their prospective achievements in the field of MCQ production. The texts are ordered as in the evaluation in ascending order in terms of

Table 2
Evaluation syllabus.

Evaluation Metrics	Score 0	Score 1	Scores 2-3	Scores 4-5
Quality of the question	Not understandable or not connected to the text	Not well formulated or has major grammatical errors	Not too relevant to the topic or too obvious	Relevant to the topic and can be answered logically
Quality of the right answer	Incorrect or not related to the question	Partially incorrect	Correct but not directly answering the question	Correct and directly answering the question
Quality of the distractors	No distractors generated	Contain grammatical errors or does not make sense	Not relevant, confusing, or reveal the right answer	Useful, relevant, seems plausible, and well-crafted

CEFR (Common European Framework of Reference) levels: A2, B2, C1, and C2, where A is beginner, B is intermediate, and C is advanced level (Tracktest, 2023). The scores were rated on a 0–5 scale, with 5 being the highest. The total average of all tools for every text is presented in the final row of the table. This total average allows us to have a clearer idea of the performance of each tool compared to the others.

4.1. Quality and reliability of the questions

The Table 3 below presents the details of the results of the collaborative evaluation concerning the first primary metric. The participants rated every tool in accordance with the evaluation syllabus provided and completed this process for the four texts.

The data presented in Table 3 represents the performance of each tool per text in terms of the quality of questions generated. Based on this, we were able to conduct a detailed comparative analysis of these tools in terms of the quality of questions generated.

In addition to the performance indicated by the means, the reliability of the given tool is also a significant factor. If the performance of a given tool is measured by the mean and its reliability by the standard deviation, taking all texts into account, the following results are obtained.

Going through Table 4, we noticed that QuizGecko performed moderately throughout the four texts, with a total average score of roughly 3.75. In general, QuizGecko had an average score lower than the total average for all four texts. The questions created by the tool were still of lower quality compared to the top performers. The standard deviation column in Table 4 shows how far the quality of the questions generated differs from the average score. The smaller the deviation, the more balanced the performance. As can be seen in Fig. 1, the low standard deviation clearly indicates that this tool has proved to be the most stable, which does not necessarily mean a good result: its performance is consistent but rather weak.

The performance of Questgen is strong in terms of question quality as it scored an average score of 4 over the four texts. Overall, its performance was usually above the total average except for the third and fourth texts, where it was slightly lower. However, we should point out

Table 3
Results for "Quality of the questions" metric.

Tools	Text 1	Text 2	Text 3	Text 4
QuizGecko	3.87	3.87	3.62	3.62
Questgen	4.37	4.25	3.87	3.5
Google Gemini	4	4.32	4.32	4.75
ChatGPT	4.12	4.25	4.37	4.62
DeepSeek	4	4.25	4.30	4.60

Table 4
Mean and standard deviation of the overall results of the generated questions.

Tools	Mean	Std dev
QuizGecko	3.75	0.14
Questgen	4.00	0.39
Google Gemini	4.35	0.31
ChatGPT	4.34	0.21
DeepSeek	4.29	0.25

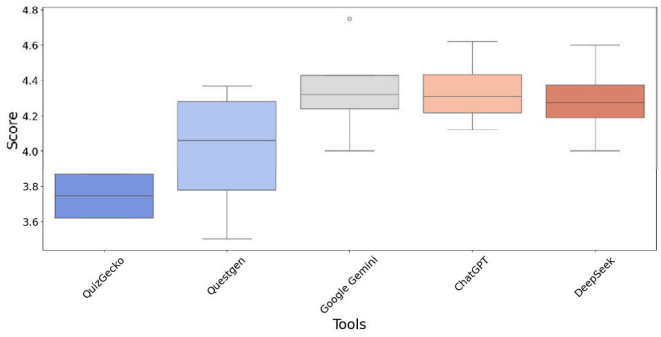

Fig. 1. Distribution of question quality scores.

that Questgen has a high standard deviation, which means that its performance is inconsistent across different texts and deteriorates in proportion to the complexity of the texts. In general, Questgen presented a good performance, most likely due to its natural language comprehension capabilities, which allow it to construct concise and contextually appropriate questions. The tool's adaptability to the nuances of individual texts is a notable advantage, but it also produces very large variations for texts of different difficulty.

The performance of Gemini concerning the quality of questions generated is impressive, with an average score of 4.35. Notably, Gemini's performance was frequently above the general average for each text. This demonstrates its capacity to handle a wide range of text difficulties and themes. However, this time, there is also a relatively high standard deviation of 0.31, so this tool is not able to provide consistently good performance in generating questions. These findings urge additional investigation into the relationship between Gemini's architecture and its ability to generate effective questions.

The high performance of ChatGPT is easily notable, with an average score of 4.34 across the four texts. The tool's scores have always been above the total average for every text, which is exceptional as it reflects its high performance on any text complexity or topic. The tool not only performs well on the quality test but also has a very good standard deviation compared to the others (0.21). These results are due to the tool's strong NLP architecture and natural language comprehension abilities. The program constantly generates meaningful, structured questions that are suited to the particular context of each text. This excellent performance across texts demonstrates ChatGPT's versatility and capacity to preserve a high level of question quality. This led to the assumption that there may be a link between this tool's performance and its NLP-driven strategy.

DeepSeek's performance was good overall, especially with the last text which reflects its ability to adapt to the complexity of language and topic discussed. It was slightly below the average in the first text but performed nearly as well as ChatGPT in the other texts, which showcased its huge potential mainly as it is still a new tool in the market. In addition, the calculated standard deviation is also quite good (0.25), which suggests that it produces relatively consistent results even for different texts.

The means clearly show that in this category, the large language models, Gemini, ChatGPT, and DeepSeek, perform significantly better

than the purpose-built quiz generators (approximately 4.3 versus 4.00 and 3.74). QuizGecko stands out, with a rather poor score of 3.74. Questgen and Google Gemini have a higher deviation (above 0.3) – these tools cannot be said to be consistent, while DeepSeek and ChatGPT perform more stably. Based on this, it can be stated that DeepSeek and ChatGPT are the ones that are most capable of providing consistently good performance when generating questions. Based on our assumption the reason for the poorer performance of purpose-built quiz generators is that they do not always use state-of-the-art generative AI models like general purpose AI chatbots.

4.2. Quality and reliability of the right answers

Table 5 below presents the details of the results of the collaborative evaluation concerning the second primary metric.

If we calculate the means and standard deviations for each text for the generated right answers, we get the following table (See Table 6).

The data presented in Table 6 represents the average performance and variance of each tool concerning the quality of the right answer generated. After analyzing the data gathered and the average scores presented in this table, we were able to conduct a detailed comparative analysis of these tools in terms of the quality of the right answer generated. Visually representing the data, it can be seen in Fig. 2 that ChatGPT and DeepSeek perform by far the best in generating the right answers.

Table 5
Results for "Quality of the right answer" metric.

Tools	Text 1	Text 2	Text 3	Text 4
QuizGecko	3.75	4.75	4.37	2.87
Questgen	4.5	4.4	4.75	3.92
Google Gemini	3.9	4.5	4.25	4.3
ChatGPT	4	4	4	4.5
DeepSeek	4.5	4.5	4.20	4.45

Table 6
Mean and standard deviation of the overall results of the right answers.

Tools	Mean	Std dev
QuizGecko	3.94	0.82
Questgen	4.39	0.35
Google Gemini	4.24	0.25
ChatGPT	4.13	0.25
DeepSeek	4.41	0.14

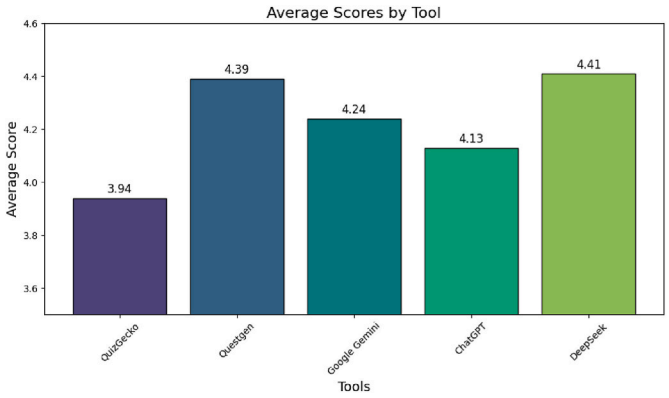

Fig. 2. Right answer quality score by tools.

QuizGecko reflected a variable performance with scores ranging from 2.87 to 4.75, indicating its inability to consistently generate high-quality answers. This tool scored the highest in text 2, and this may be due to its capability to efficiently answer factual questions. However, the tool's performance was remarkably lower than the average for text 4, and this may be due to its difficulties in offering accurate responses to questions about complex topics. Another assumption is that QuizGecko might be relying on pre-defined templates that do not always fit individual text characteristics. In terms of stability, QuizGecko's performance is the most variable, with this tool having the highest standard deviation (~0.82). Due to the results it provides, manual review is required to generate correct answers, as the quality varies within a fairly wide interval.

Questgen showed excellent performance across the four texts. Its score was always above the total average which means that it generated good answers for the questions compared to the other tools. The tool also presented a lower performance for the fourth text compared to the results of the three previous ones, and this can be due to the complexity of the text. However, the tool was generally successful at correctly answering the questions. The standard deviation is relatively high, indicating that the performance is not uniform, but it is still the best after DeepSeek. Questgen's good performance can be attributed to its NLP-driven strategy, which enables it to perceive the contextual nuances of each text and provide high-quality correct answers. This led us to assume that there is a correlation between the NLP approach of Questgen and the high-quality answers it provided in the MCQs.

Gemini's performance was good and consistent throughout the texts ranging from 3.9 to 4.5. Although its score in the first text was slightly lower than the total average, its overall performance in the subsequent texts was notably superior. The standard deviation is 0.25, which is acceptable considering all the tools examined. Gemini's integration of NLP-driven techniques may be credited with this result. These techniques vary from tokenization to semantic analysis and others. The program is capable of producing high-quality answers that adhere to the set requirements for each text.

ChatGPT's results showed that it also achieved high results consistently over the four texts by having an average of 4.13. Even though its average score was not always higher than the total average for all texts, it kept a good performance overall. The standard deviation is the same as that obtained by Gemini: it could be better, but considering the rivals, it is adequate. A possible assumption is that there is a significant relation between its strong language processing capabilities and its capacity to provide high-quality answers. The fact that ChatGPT consistently performed well in all texts demonstrates its language understanding and versatility to provide contextually appropriate and correct answers.

DeepSeek demonstrated strong performance across all four texts, consistently generating correct answers above the overall average. Additionally, its total average score was the highest among the evaluated tools which shows its high capability of generating high-quality answers in the MCQs. These results highlight DeepSeek's potential, particularly as further enhancements are implemented. We can state that in terms of correct answers, DeepSeek's performance is the most stable, with low variance (0.14), and produces exceptionally good results in performance (4.41). The winner of this category is undoubtedly DeepSeek.

4.3. Quality and reliability of the distractors

Table 7 presents the details of the results of the collaborative evaluation concerning the third primary metric.

The data presented in Table 7 represent the average performance of each tool concerning the quality of the distractors generated. After analyzing the data gathered, the mean scores, and standard deviation values presented in Table 8, we were able to conduct a detailed comparative analysis of those tools in terms of the quality and reliability of the distractors generated.

Table 7
Results for "Quality of the distractors" metric.

Tools	Text 1	Text 2	Text 3	Text 4
QuizGecko	4.5	4.12	4	3
Questgen	4.37	4.25	4.25	3.5
Google Gemini	4.21	4.37	4	4
ChatGPT	4.25	4.25	4.25	4.62
DeepSeek	4.20	4.20	4.25	4.35

Table 8
Mean and standard deviation of the overall results of the Distractors.

Tools	Mean	Std dev
QuizGecko	3.91	0.64
Questgen	4.09	0.40
Google Gemini	4.15	0.18
ChatGPT	4.34	0.19
DeepSeek	4.25	0.07

Having an average score of 3.91, QuizGecko performed well in this area. Across all four texts, it consistently created high-quality distractors. The scores ranged from 4 to 4.5 except for the last text, which might be because the text discusses a complex topic containing multiple technical terms. However, the participants in the evaluation thought that the wrong answer selections were relevant, well-crafted, and appropriately difficult. These otherwise useable results deteriorated as the text became more complex. We can state that QuizGecko is not capable of stable, reliable performance, with a standard deviation of 0.64, the least stable of all tools.

Questgen has also shown an excellent performance over the four texts concerning this metric with an average score of 4.09. It was capable of keeping a score greater than the total average which shows its good capacities regarding the generation of high-quality distractors compared to the other tools. However, we noticed that the tool generated lower-quality distractors for the last text, and this may be because of the complexity of the texts and terms used. The produced distractors were also relevant to the questions and provided a reasonable amount of challenge. Regarding the consistent nature of the results, the standard deviation of 0.4 is quite high compared to the LLM models examined, indicating that it is strongly dependent on the complexity of the text.

From the data presented in Table 8, Gemini has performed very well while generating distractors with an average score of 4.15. Even though its different scores across the texts were not always above the total average, its performance was steadily good. There is a difference in the quality of distractors generated for the last two texts compared to the first ones, as the scores are lower, this may reflect the difficulty for Gemini to keep up the same performance for more complex texts using technical terms for example. However, it performs exceptionally well in terms of standard deviation ($\sigma = 0.18$), so high performance is coupled with reliability.

Concerning the metric of the quality of distractors generated, ChatGPT stands out compared to the other tools as it has an average score of 4.34 across the four texts. It has consistently generated high-quality distractors that always scored above the total average which reflects the tool's good performance in terms of generating useful and relevant distractors. It also holds its own when examining the deviation from the mean, with a value of 0.19 just a hair higher than the value given by Gemini.

For every text, DeepSeek consistently generated good distractors, scoring from 4.20 to 4.35, for an average of 4.25l. DeepSeek's performance was competitive with other tools like Questgen and Google Gemini, even though ChatGPT had a better overall performance with an average of 4.34. These findings show that DeepSeek can consistently produce plausible, challenging, and contextually relevant distractors, even when applied to texts with different complexity levels. In terms of standard deviation, DeepSeek gave the best result by far; the value of

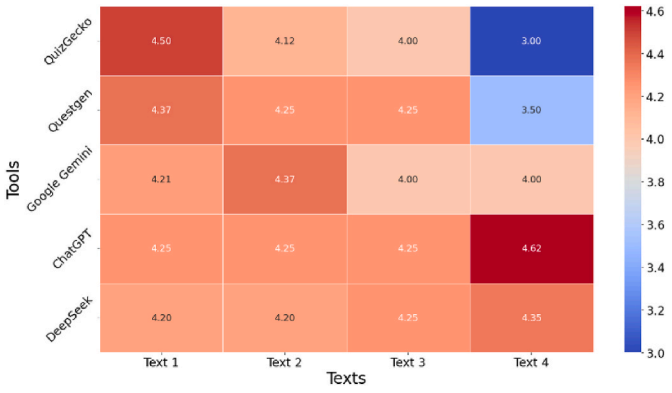

Fig. 3. Heatmap of distractor quality scores.

0.07 indicates almost negligible, consistent, and reliable performance. These results highlight the tool's potential as a reliable software for automatic MCQ generation.

Although the heatmap in Fig. 3 provides useful visual cues, on its own, it does not clearly indicate that each tool specializes in different types of texts, at least not strongly or consistently. While there are some minor differences, overall, the tools perform similarly across the texts, with only minor fluctuations that could be due to small contextual differences rather than specialization. For instance, while ChatGPT shows notably higher performance on Text 4 (4.62), and QuizGecko drops significantly on the same text (3.00), these patterns are not consistently mirrored across other texts or tools. This suggests that such variations are likely due to text-specific complexity or tool limitations rather than evidence of specialization.

4.4. *Aggregated results of the tools*

In this subsection, the results in the different categories were summarized to see if there was any tool that had outperformed the others in all respects. These results directly address RQ3 by systematically comparing the strengths and weaknesses of general-purpose LLMs against domain-specific MCQ generators. Additionally, we wanted to clarify whether the LLMs performed better overall than the target software, QuizGecko and Questgen.

For this purpose, we created a table in which we aggregated the subperformances and variances obtained in each category and thus compared the five participants with each other. The result of the aggregation is shown in Table 9.

The data was analyzed to evaluate the overall performance of the five AI-based question-generation tools on three key quality indicators: the quality of the questions generated, the quality of the correct answers, and the quality of the distractors (see Fig. 4). To gain a deeper understanding of the results, we used statistical computations, various data visualization techniques, cluster analysis, and network analysis, which allowed us to compare the performance of the tools and identify similarities and differences between them.

Looking at the means and standard deviations, it can be seen that Questgen and DeepSeek tools show the highest overall performance (4.32 and 4.29) (see Table 9). Questgen achieved a slightly higher average score, while DeepSeek produced more stable results with lower

Table 9
Aggregated statistical measures.

Tools	Mean	Std dev
QuizGecko	4,04	0.32
Questgen	4.32	0.27
Google Gemini	4.10	0.17
ChatGPT	4.23	0.17
DeepSeek	4.29	0.18

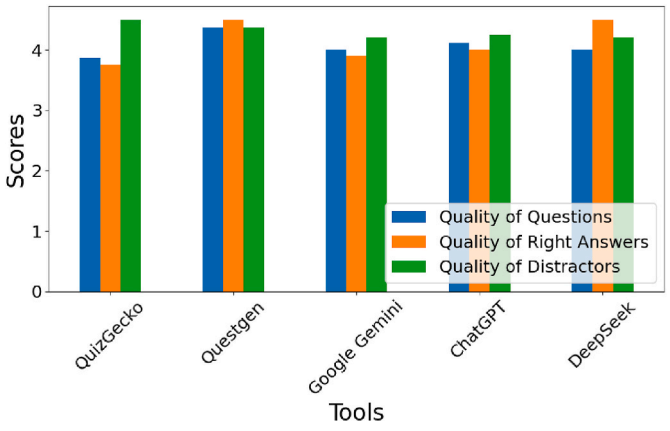

Fig. 4. Performance of each tool.

standard deviations. ChatGPT also showed consistently high scores, especially in the quality of correct answers. In contrast, QuizGecko's performance fluctuated significantly, especially for correct answers, where its scores showed the lowest minimum values, indicating that its generated answers are less reliable.

When examining the correlation between the different quality indicators, it was found that there is a strong positive relationship between the quality of the questions and the quality of the right answers (see Fig. 5). This suggests that tools that generate higher-quality questions tend to produce higher-quality correct answers. The quality of distractors, on the other hand, showed a weaker relationship with the other two variables, suggesting that generating relevant but incorrect answers may be a particular challenge for the tools examined.

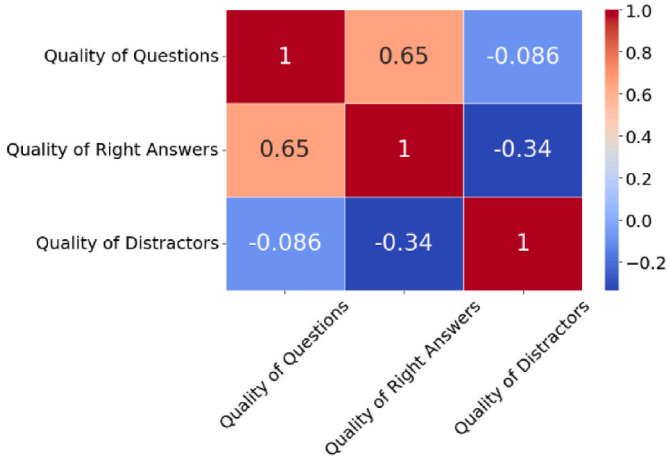

Fig. 5. Correlation heatmap of quality metrics.

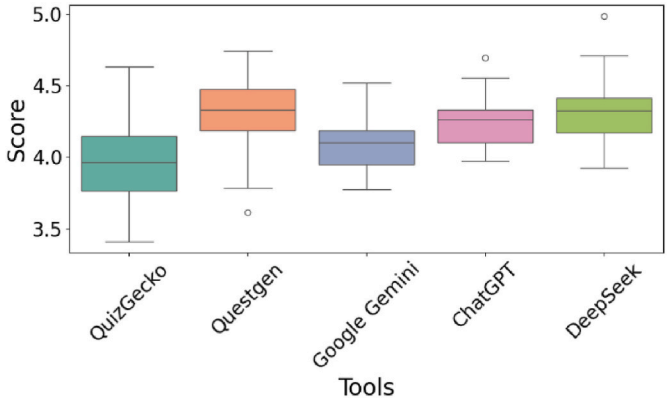

Fig. 6. Distribution of scores.

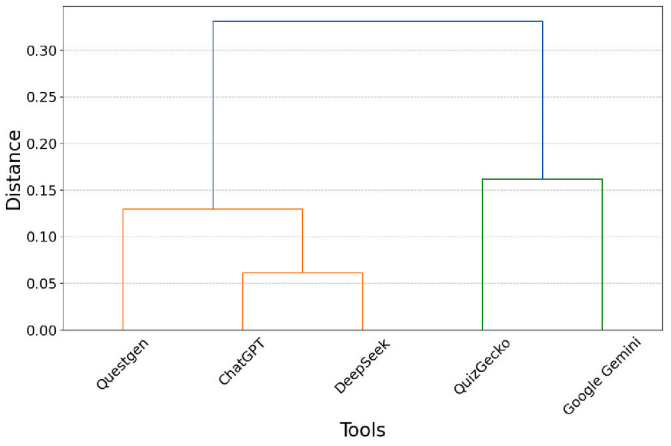

Fig. 7. Dendrogram of the tools examined.

As shown in Fig. 6, the box plots for Questgen and DeepSeek typically display a higher median value, indicating that these tools produce consistent and high-quality outputs. ChatGPT and Gemini tend to show a moderately high value. This suggests that while their overall performance is strong, there is a bit more variability in their scores. QuizGecko's box plot is positioned lower on the Y-axis, indicating a lower median score. This implies that QuizGecko is not only producing lower quality outputs on average, but its performance is also more inconsistent.

In the case of the dendrogram (Fig. 7), the wood structure shows the similarity between the subjects studied. If two tools are quickly linked, they are very similar. If they are joined later, the difference is greater. The figure shows that DeepSeek and ChatGPT performed very similarly, Questgen is slightly different, but even so, it can be seen that these tools belong to the same category. The remaining tools are more distinct; most notably QuizGecko, which is by far the weakest performer. Gemini performed better, but there is still a significant difference. Accordingly, the dendrogram can clearly identify two clusters: 1) Questgen, ChatGPT, and DeepSeek and 2) a cluster of QuizGecko and Questgen.

The result of K-means clustering, shown in Fig. 8, produced visually distinct groups based on the performance and deviation of the tools and also confirmed those already indicated by the dendrogram. One cluster contains the more stable, predictable models with lower deviation, while the other cluster contains the tools with more volatile performance. The cluster midpoints are a good representation of the main characteristics of the two groups and demonstrate that there are indeed

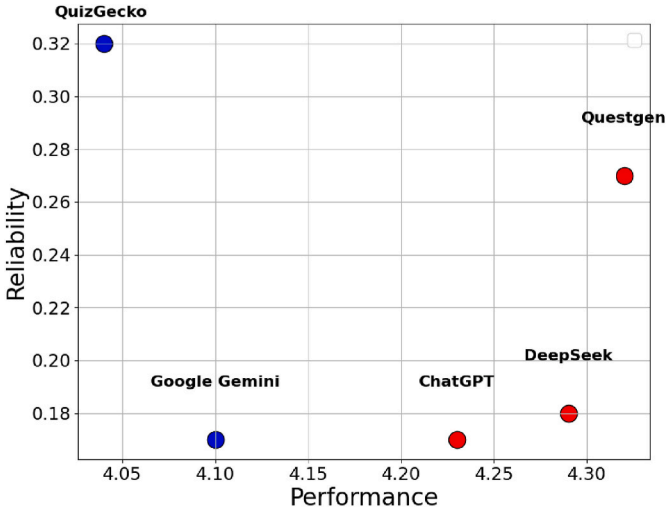

Fig. 8. K-Means clustering of tools.

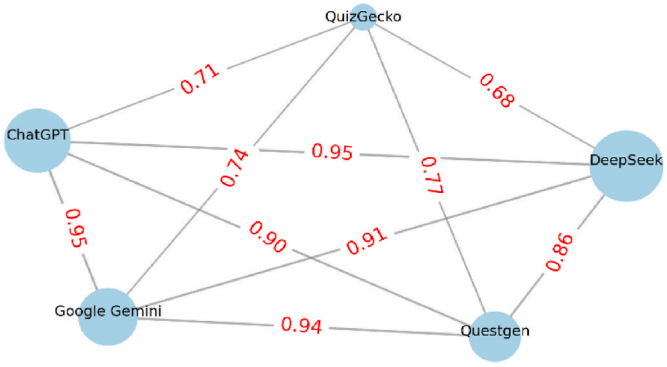

Fig. 9. Network analysis.

two different trends in terms of model performance and reliability. The fact that the location of the points is so clearly distinct shows that the bivariate approach is effective in highlighting differences between tools. As a result, models with higher variance may show more significant fluctuations in results, while models with lower variance show better consistency.

In this network analysis in Fig. 9, each node represents one of the five tools (QuizGecko, Questgen, Google Gemini, ChatGPT, and DeepSeek). The average scores aggregated from the different quality metrics are used as the basis for determining the similarity between the tools. In the visualization, tools with similar performance have a high similarity score and thus an edge with a higher number (e.g. ChatGPT, DeepSeek, and Google Gemini are therefore strongly related, while QuizGecko is only weakly related to the other tools due to its lower average and higher variance). The size of the nodes also indicates the level of performance. From the network analysis, it is clear that ChatGPT and DeepSeek form a strong cluster, highlighting their excellent and consistent performance, while QuizGecko has a much weaker connection with the others due to its relatively weaker performance.

5. Secondary metrics

For the evaluation of the tools compared in terms of the secondary metrics, we conducted an individual assessment by using each program several times on different texts. As mentioned previously, these metrics include the ease of use of the tools, their speed of generation of the MCQs, and their individual limitations.

5.1. Evaluation syllabus

Concerning the Secondary metrics, we created the following standard framework to effectively assess the tools and further address RQ2. The same scoring system used for the primary metrics evaluation is also used in this section for the secondary metrics stated in section 3.3. Table 10 provides the details of the evaluation system used.

5.2. Results

Table 11 presents the summary of our secondary metrics evaluation. The scores presented in the table represent the average rating across the four different texts to have an aggregated view of each tool's performance.

Table 11
Results of secondary metrics evaluation.

Tools	Ease of use	Speed of Generation	Individual Limitations	Customization Flexibility
QuizGecko	5	1	0	2
Questgen	5	5	3	4
Google Gemini	5	5	5	5
ChatGPT	5	5	5	5
DeepSeek	5	5	5	5

Regarding the ease of use of the tools, we noticed while using QuizGecko and Questgen that they offered the ability to choose the number of questions to generate, the language, and the difficulty of the MCQ generated. This is an efficient way to help instructors customize their quizzes depending on their needs, especially since the tools' interfaces were easy to navigate in and choose the needed options. To create MCQs using these tools, the user needs to enter the text, choose the desired options, and submit. Concerning Google Gemini, ChatGPT, and DeepSeek, as they are chatbots, the user only needs to enter an inquiry providing the details of the MCQ generation paired with the corresponding text. This makes these tools highly flexible in terms of customization, as users can specify parameters such as the number, type, or difficulty of the questions using natural language prompts. This flexibility is reflected in the Customization Flexibility metric included in Table 11.

For the speed of generation, most of the tools did not showcase any latency while generating the MCQs. However, for QuizGecko, it took a notably long time to analyze the texts and generate the MCQs which can negatively affect the user experience.

In terms of the individual limitations of each tool, we noticed while using QuizGecko that it imposes multiple constraints for the user in case of using the free version of the program such as: not exceeding 1000 characters of text, not being able to generate more than 5 questions MCQs, not being able to use the program for more than 2 texts. These conditions can cause problems when using the free version of the software as it does not give a large space to experience the platform. The user needs to pay for the software if they want to use it frequently. Questgen also offers two plans, the free and the paid ones. However, the user can use the free plan more freely than in the case of QuizGecko, as they can enter up to 5000 characters and run the program 13 times before the need for payment. This allows the user to have a clearer idea about the performance of the tool before deciding to have the paid plan. Concerning ChatGPT, it presents a free and a paid version of the chatbot, however, there are no restrictions on usage for the free plan. The difference between the two versions is that the free one uses GPT-3.5 while the paid one uses the GPT-4 model which is more advanced and presents better results. For Google Gemini, it offers a free version that is effective for tasks such as question generation, and a paid version that is useful for more complex tasks. Finally, DeepSeek is free and can be used without any restrictions as all its features are accessible without a paid tier in Q1 2025.

6. Challenges and recommendations

Our comparative analysis of AI-based MCQ generators identified

Table 10
Secondary metrics evaluation syllabus.

Evaluation Metrics	Score 0	Score 1	Scores 2-3	Scores 4-5
Ease-of-Use	Not useable, very complicated interface	Poor usability with significant issues	Minor usability issues, workable	Highly intuitive and user-friendly
Speed of Generation	Unacceptably slow, long delays	Noticeable slowness with frequent delays	Acceptable speed with occasional lag	Very fast and responsive
Individual Limitations	Extremely restrictive free version	Significant restrictions	Moderate restrictions	Minimal or no restrictions

three important issues that define the level of automated question generation today. These findings outlined in this section answer RQ4 by highlighting the main issues in AI-based MCQ generation tools and suggesting actionable recommendations for quiz creators. One of the most notable findings was the stronger performance of LLMs such as ChatGPT, Gemini, and DeepSeek in handling complex texts. The study showed that these general-purpose LLMs continuously maintained or even enhanced their performance when text complexity rose, especially in the C1 and C2 level texts. This pattern implies that these models have a clear advantage in comprehending complex language and topics because of their extensive training and excellent language understanding skills.

On the other hand, QuizGecko and Questgen, two specialized MCQ generators, demonstrated decreasing performance as text complexity increased. These tools showcased a higher performance with simpler texts, but they became less useful when dealing with complicated material. This performance disparity highlights a core issue facing the industry: the trade-off between specialized functionality and adaptability.

Furthermore, there are major operational issues due to the practical constraints of specialized tools. Significant usage restrictions, such as character limits that impair the ability to analyze lengthy messages, are frequently imposed by the free versions of these programs. Moreover, we found that the processing speed of the various tools differed significantly, which may affect their usefulness in educational contexts where time efficiency is essential.

Even while all of the tools showed proficiency in producing basic MCQs, it is still difficult to maintain uniform quality across various topic areas and degrees of difficulty. This is especially true when it comes to the creation of distractors, where the technical aspect of the material may greatly affect the quality.

Based on these patterns, we recommend a strategic approach to tool selection and implementation. Our results support the use of general-purpose LLMs as the more reliable choice for complex materials. On the other hand, specialized tools may be more suited for basic subjects where their extra formatting and customization options are useful. Businesses should consider adopting a hybrid strategy that makes use of various technologies according to certain use cases and the complexity of the material.

Concerning tool developers, our results suggest prioritizing the improvement of adaptable features in specialized tools while preserving their distinct formatting and customization benefits. Reducing resource limitations and increasing processing effectiveness without sacrificing output quality should be the main goals of future developments.

These results and recommendations offer a foundation for the present deployment as well as future advancement of AI-based MCQ generators, especially in corporate digital learning settings where the creation of effective, high-quality assessments is essential.

7. Limitations and future work

One of the challenges of this study is the size of the dataset used for benchmarking. Although the texts selected cover a range of CEFR levels (A2 to C2) and different topics, their small number restricts the generalizability of the findings. Having a small sample, subtle performance differences between the tools can be exaggerated, and the statistical power of the comparative tests is reduced. Additionally, the size of the dataset may not sufficiently represent the large diversity of real-world texts that teachers use. These can include informal reading passages, dialogues, scientific articles, policy documents, and exam texts.

The decision to keep the dataset small was intentional and backed by the need for manual, human evaluation. Since we used expert ratings to assess question quality, correct answers, and distractors, a larger dataset would have made the process time-consuming and difficult to manage within the available resources. While human evaluation provided valuable insights and ensured educational relevance, it also introduced a degree of subjectivity. Different evaluators may have slightly different expectations or standards. Even with a shared rubric, reviewers may interpret questions or distractors differently. This limits the scalability of the evaluation method.

The language scope of the study is another limitation. All texts used in the dataset were in English. Many classrooms are multilingual, especially in global or diverse learning environments. For this reason, future studies should include texts in other languages to reflect real classroom contexts. Additionally, the dataset could be improved by adding more subject-specific content (e.g. biology, law, history …). This is because MCQ generation in technical subjects is more demanding and can expose other strengths and weaknesses of the tools.

Future studies could complement human evaluation with automated measures. Readability metrics like Flesch-Kincaid, SMOG, or Dale-Chall can give quick insights into text complexity. Semantic similarity tools like BERTScore, BLEU, or ROUGE can be used to compare generated items with expected outputs. For distractors, lexical distance tools can help measure the similarity or difference objectively.

8. Conclusions

Based on the results presented in the comparative analysis section, the primary metrics yielded some noteworthy conclusions. ChatGPT, DeepSeek, and Google Gemini outperformed in these primary criteria, owing largely to their good NLP-driven techniques and the huge number and diversity of the training data they use. Questgen and QuizGecko were closely behind, providing medium to high performance, although this varied depending on the assessment metric and text utilized. These findings are in line with prior studies such as Firdaus et al. (2024), which also highlighted the strengths of LLMs in question generation, especially when dealing with context-rich input. However, while Firdaus et al. focused primarily on LLM capabilities, our comparative analysis offers a broader perspective, incorporating both general-purpose and domain-specific tools. This supports existing claims but adds nuance regarding consistency and distractors' quality.

When the tools were evaluated using secondary metrics, a new set of observations appeared. In terms of usability, QuizGecko and Questgen provided user-friendly interfaces with the ability to alter quiz criteria, making them ideal for instructors looking for tailored quizzes. ChatGPT, DeepSeek, and Google Gemini provided personalization depending on user inquiries, with no predefined constraints. In terms of the speed of generation, most tools worked well, except QuizGecko, which had a noticeably longer processing time, potentially affecting the user experience. In this part, the limitations of each program were noted. Both QuizGecko and Questgen have free and premium plans, although the free versions include use restrictions such as character count and the number of MCQs created. ChatGPT, DeepSeek, and Google Gemini offer both free and premium versions, with no substantial use limits on the free plan.

As far as performance evaluation is concerned, the metrics show an even more mixed picture In the competition within the category, DeepSeek beats ChatGPT and Gemini. Similarly, Questgen performs better than the underperforming QuizGecko. Our results also indicate that there is no significant difference between general-purpose AI and purpose-built AI in the context of MCQ generation. However, there are early signs that LLMs may handle more complex texts reliably, as seen in their performance on higher-level input like Text 4. This observation should be further investigated using a larger and more diverse dataset.

The collaborative evaluation method used in this study was designed to ensure consistency and fairness, with ten experts rating the outputs independently using a predefined rubric. While human evaluation

involves some subjectivity, the structure and repeated scoring across tools and texts strengthen the reliability of the results.

These findings contribute to a broader understanding of current AI-based MCQ generators and can guide both instructors and institutions in selecting tools that align with their needs. While the results are based on a limited dataset, the variety of topics and CEFR levels used offers a degree of generalizability to real-world educational contexts. As demand for scalable, intelligent assessment tools increases, further research should extend these comparisons to other domains, languages, and emerging AI models to build a more comprehensive view of the field.

CRediT authorship contribution statement

Asmae Azzi: Writing – original draft, Validation, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Ferenc Erdős:** Writing – original draft, Validation, Supervision, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Richárd Németh:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Formal analysis, Data curation. **Vijayakumar Varadarajan:** Writing – review & editing, Supervision. **Stephen Afrifa:** Writing – review & editing.

Statements on ethics and open data

Informed consent was obtained from all participants, and their privacy rights were strictly observed. Prior to their participation, all participants were informed that the collected data would be used for research purposes and would be handled anonymously. Detailed information about the study was provided to participants before beginning the survey, reassuring them of their right to voluntary participation and withdrawal. The data can be obtained by sending an e-mail request to the corresponding author.

Funding

This work did not receive any financial support.

Declaration of competing interest

The author is an Editorial Board Member/Editor-in-Chief/Associate Editor/Guest Editor for this journal and was not involved in the editorial review or the decision to publish this article.

References

Agarwal, M., Sharma, P., & Goswami, A. (2023). Analysing the applicability of ChatGPT, bard, and bing to generate reasoning-based multiple-choice questions in medical physiology. *Cureus, 15/2023*. https://doi.org/10.7759/cureus.40977

Ahmed, I., Roy, A., Kajol, M. A., & Reza, M. R. (2023). ChatGPT vs. Bard: A comparative study. *Engineering Reports, 6*(11). https://doi.org/10.36227/techrxiv.23536290

Azaria, A., et al. (2023). ChatGPT is a remarkable tool – for experts. *Data Intelligence, 6* (1). https://doi.org/10.1162/dint_a_00235

Barbieri, C. A., & Rodrigues, J. M. (2025). Leveraging cognitive load theory to support students with mathematics difficulty. *Educational Psychologist*. https://doi.org/10.1080/00461520.2025.2486138

Bhandari, S., Liu, Y., Kwak, Y., & Pardos, Z. A. (2024). Evaluating the psychometric properties of ChatGPT-generated questions. *Computers and Education: Artificial Intelligence, 7*, Article 100284. https://doi.org/10.1016/j.caeai.2024.100284

Bhratha, A., et al. (2024). Comparing the performance of ChatGPT-4 and medical students on MCQs at varied levels of Bloom's taxonomy. *Advances in Medical Education and Practice, 15*. https://doi.org/10.2147/amep.s457408 (2024).

Bi, X., et al. (2024). DeepSeek LLM scaling open-source Language Models with longtermism. https://www.researchgate.net/publication/379694907_DeepSeek_LLM_Scaling_Open-Source_Language_Models_with_Longtermism. (Accessed 16 February 2025).

Bitew, S. K., Hadifar, A., Sterckx, L., Deleu, J., Develder, C., & Demeester, T. (2024). Learning to reuse distractors to support multiple-choice question generation in education. *IEEE Transactions on Learning Technologies, 17*, 375–390.

Bolgova, O., Shypilova, I., Sankova, L., & Mavrych, V. (2023). How well did ChatGPT perform in answering questions on different topics in gross anatomy? *European Journal of Medical and Health Sciences, 5/2023*. https://doi.org/10.24018/ejmed.2023.5.6.1989

Chowdhery, A., Narang, S., & Devlin, J. (2022). PaLM: Scaling language modeling with pathways. In *Long beach: Proceedings of the 36th international conference on machine learning*.

Collins, J. (2006). Education techniques for lifelong learning: Writing multiple-choice questions for continuing medical education activities and self-assessment modules. *RadioGraphics: A Review Publication of the Radiological Society of North America, Inc, 26*(2), 543–551. https://doi.org/10.1148/RG.262055145

Deming, C., Dekkati, S., & Desamsetti, H. (2018). Exploratory data analysis and visualization for business analytics. *Asian Journal of Applied Science and Engineering, 7* (1). https://doi.org/10.18034/ajase.v7i1.53

Ebenbeck, N., Bastian, M., Mühling, A., & Gebhardt, M. (2024). Duration versus accuracy—what matters for computerised adaptive testing in schools? *Journal of Computer Assisted Learning, 40*, 3443–3453. https://doi.org/10.1111/jcal.13074

Firdaus, T., Sholeha, S., Jannah, M., & Setiawan, A. (2024). *Comparison of ChatGPT and Gemini AI in answering higher-order thinking skill biology questions: Accuracy and evaluation, Kopi alinea's Lab*.

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines, 30*(2). https://doi.org/10.2139/ssrn.3827044

Gaur, A. (2025). DeepSeek vs. OpenAI: What is DeepSeek? What does it do?. https://mindflow.io/blog/deepseek-vs-openai-what-is-deepseek-what-does-deepseek-do. (Accessed 16 February 2025).

gouthamg, r. g. (2023). Question generation using state-of-the-art Natural Language Processing algorithms. https://github.com/ramsrigouthamg/Questgen.ai#nlp-models-used. (Accessed 3 November 2023).

Haataja, E., Tolvanen, A., Vilppu, H., Kallio, M., Peltonen, J., & Metsäpelto, R.-L. (2023). Measuring higher-order cognitive skills with multiple-choice questions—potentials and pitfalls of Finnish teacher-education entrance. *Teaching and Teacher Education, 122*, Article 103943. https://doi.org/10.1016/j.tate.2022.103943

Haladyna, T. M., & Downing, S. M. (2004). *Developing and validating test items*. Mahwah: Lawrence Erlbaum Associates Publishers.

Herrmann-Werner, A., Festl-Wietek, T., Holderried, F., Herschbach, L., & colleagues. (2024). Assessing ChatGPT's mastery of Bloom's taxonomy using psychosomatic medicine exam questions: Mixed-methods study. *Journal of Medical Internet Research, 26*, Article e52113. https://doi.org/10.2196/52113

Khosravi, H., Buckingham Shum, S., Chen, G., Conati, C., Tsai, Y.-S., Kay, J., … Gašević, D. (2022). Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence, 3*, Article 100074. https://doi.org/10.1016/j.caeai.2022.100074

Komorowski, M., Marshall, D. C., Salciccioli, J. D., & Crutain, Y. (2016). Exploratory data analysis. In MIT Critical Data (Ed.), *Secondary analysis of electronic health records*. Cham: Springer.

Kumar, A., Nayak, A., Shenoy, M., Goyal, S., & Chaitanya. (2023). A novel approach to generate distractors for Multiple Choice Questions. *Expert Systems with Applications, 225/2023*. https://doi.org/10.1016/j.eswa.2023.120022

Kurdi, M., Haffari, G., & Rasooli, M. S. (2020). Automatic multiple choice question generation from text: A survey. In *IEEE transactions on learning technologies* (pp. 14–25). IEEE.

Lewis, J. R. (2014). Usability: Lessons learned … and yet to Be learned. *International Journal of Human-Computer Interaction, 30*(9), 663–684. https://doi.org/10.1080/10447318.2014.930311 (10).

Liu, Y., Bhandari, S., & Pardos, Z. A. (2025). Leveraging LLM respondents for item evaluation: A psychometric analysis. *British Journal of Educational Technology, 56*, 1028–1052. https://doi.org/10.1111/bjet.13570

Lu, S., Wang, M., Liang, S., Lin, J., & Wang, Z. (2020). Hardware accelerator for multi-head attention and position-wise feed-forward in the transformer. In *2020 IEE 33rd international system-on-chip conference (COCC)*. USA: Las Vegas. https://doi.org/10.1109/socc49529.2020.9524802.

Newton, P. M., & Xiromeriti, M. (2024). ChatGPT performance on multiple-choice question examinations in higher education: A pragmatic scoping review. *Assessment & Evaluation in Higher Education, 49*(6), 781–798. https://doi.org/10.1080/02602938.2023.2299059

Noda, R., Tanabe, K., Ichikawa, D., & Shibagaki, Y. (2025). GPT-4's performance in supporting physician decision-making in nephrology multiple-choice questions. *Scientific Reports, 15*, Article 15439. https://doi.org/10.1038/s41598-025-99774-3

Prabhu, G., & Prabhu, R. (2023). ChatGPT performance on scenario-based multiple-choice questions (MCQs) in medical physiology. *Research Square*. https://doi.org/10.21203/rs.3.rs-3460822/v1

Questgen. (2023). Questgen. https://app.questgen.ai/. (Accessed 3 November 2023).

QuizGecko. (2023). QuizGecko. https://quizgecko.com/. (Accessed 3 November 2023).

Raffel, C., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. In *Vancouver: Proceedings of the 33rd international conference on neural information processing systems*.

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. In *Online: Proceedings of the 58th annual meeting of the association for computational linguistics*. https://doi.org/10.1162/tacl_a_00349

Saeidnia, H. R. (2023). Welcome to the Gemini era: Google DeepMind and the information industry. *Library Hi Tech News*. https://doi.org/10.1108/lhtn-12-2023-0214

Sallam, M., et al. (2024). The performance of OpenAI ChatGPT-4 and Google Gemini in virology multiple-choice questions: A comparative analysis of English and Arabic responses. *BMC Research Notes, 17*(1). https://doi.org/10.1186/s13104-024-06920-7

Sha, H. (2022). Understanding dichotomous questions: Examples, benefis & alternatives. https://surveypoint.ai/blog/2022/12/01/understanding-a-dichotomous-questi

ons-examples-benefits-alternatives/#What_are_Dichotomous_Questions. (Accessed 1 May 2024).
Tracktest. (2023). English language levels (CEFR). https://tracktest.eu/english-levels-cefr/. (Accessed 9 March 2025).
Vaswani, A., et al. (2017). Attention is all you need. In *31st conference on neural information processing systems, long beach, CA, USA*.
Wolfe, J. H. (1976). Automatic question generation from text - an aid to independent study. In *Technical symposium on computer science education*. https://doi.org/10.1145/952989.803459
Yang, C., Li, J., Zhao, W., & Luo, L. (2023). Do practice tests (quizzes) reduce or provoke test anxiety? A meta-analytic review. *Educational Psychology Review, 35*, 87. https://doi.org/10.1007/s10648-023-09801-w
Ye, J., & Pandarinath, C. (2021). Representation learning for neural population activity with Neural Data Transformers. *Neurons, Behavior, Data analysis, and Theory, 5*(3). https://doi.org/10.1101/2021.01.16.426955
Zhu, Q., et al. (2024). *DeepSeek-Coder-V2: Breaking the barrier of closed-source models in code intelligence* [Online] Available at: 10.48550/arXiv.2406.11931 . (Accessed 16 February 2025).