# Using AI-powered multiple-choice question generation for self-regulated learning

## Uso de generación de preguntas de opción múltiple basada en inteligencia artificial para el aprendizaje autorregulado

Enrique Barra
Universidad Politécnica de Madrid, Madrid, Spain
enrique.barra@upm.es

Anabel Pilicita
Universidad Politécnica de Madrid, Madrid, Spain
a.pilicita@alumnos.upm.es

Javier Conde
Universidad Politécnica de Madrid, Madrid, Spain
javier.conde.diaz@upm.es

Alejandro Pozo
Universidad Politécnica de Madrid, Madrid, Spain
alejandro.pozo@upm.es

Sonsoles López-Pernas
University of Eastern Finland, Joensuu, Finland
sonsoles.lopez@uef.fi

Pedro Reviriego
Universidad Politécnica de Madrid, Madrid, Spain
pedro.reviriego@upm.es

**Abstract**
This study examines the integration of generative AI in education, specifically evaluating AI-generated multiple-choice questions (MCQs) and their role in supporting self-regulated learning (SRL). Using AIQUIZ, an open-source AI-driven platform, 325 of the 593 enrolled students (54.8%) across four computing courses (Web Technologies and Databases) used the platform and generated 38,752 MCQs over two years. An explanatory sequential mixed-methods design analysed student performance, error reports, survey insights, and expert evaluations. Results showed a 70.79% overall student performance (79.45% in Databases, 66.84% in Web Technologies). Only 0.85% of questions were flagged by students as potentially incorrect, a figure that reflects user perception rather than a verified error rate. Surveys indicated strong student acceptance, engagement, and motivation, which are vital for the forethought phase of SRL. However, error analysis of flagged items revealed recurring issues like incorrectly marked answers and flawed distractors. These findings suggest that AI-generated MCQs may support the SRL cycle by facilitating forethought, performance control, and self-reflection. While Large Language Model (LLM) tools provide scalable opportunities for practice and self-assessment, our results confirm that human validation remains essential to ensure content quality and maximize learning benefits.
**Key words:** AI-Generated Multiple-Choice Questions, Generative AI in Education, Adaptive Learning, AI in Higher Education, Large Language Models.

**Resumen**
Este estudio examina la integración de la IA generativa en la educación, evaluando específicamente las preguntas de opción múltiple (MCQs) generadas por IA y su papel en el aprendizaje autorregulado (SRL). Utilizando AIQUIZ, una plataforma de código abierto impulsada por IA, 325 de los 593 estudiantes matriculados (54,8%) en cuatro cursos de informática (Tecnologías Web y Bases de Datos) usaron la plataforma y generaron 38.752 MCQs durante dos años. Se empleó un diseño de métodos mixtos secuencial explicativo para analizar el rendimiento estudiantil, informes de errores, encuestas y evaluaciones de expertos. Los resultados mostraron un rendimiento estudiantil del 70,79% (79,45% en Bases de Datos, 66,84% en Tecnologías Web). Únicamente el 0,85% de las preguntas fue señalado por los estudiantes como potencialmente incorrecto, cifra que refleja percepción de usuario y no puede interpretarse como tasa real de error. Las encuestas indicaron gran aceptación, compromiso y motivación, factores vitales para la fase de planificación del SRL. Sin embargo, el análisis de errores reveló problemas recurrentes, como respuestas mal marcadas y distractores defectuosos. Estos hallazgos sugieren que las MCQs generadas por IA pueden apoyar eficazmente el ciclo SRL al facilitar la planificación, el control del rendimiento y la autorreflexión. Aunque los modelos de lenguaje grande (LLM) ofrecen oportunidades escalables para la autoevaluación, la validación humana sigue siendo esencial para garantizar la calidad del contenido y maximizar el aprendizaje.
**Palabras clave:** Preguntas de opción múltiple generadas por IA, IA generativa en la educación, aprendizaje adaptativo, IA en la educación superior, grandes modelos de lenguaje

## 1. Introduction

In higher education, fostering self-regulated learning (SRL) is crucial for students to actively monitor, evaluate, and adjust their learning strategies (Zimmerman, 2000). To frame our study, we adopt the social cognitive perspective of Zimmerman (Zimmerman, 2000), who defines SRL as self-generated thoughts, feelings, and actions that are planned and cyclically adapted to the attainment of personal goals. This model posits that SRL operates in a three-phase cycle. In the forethought phase, processes that precede learning efforts take place, including task analysis, goal setting, strategic planning, and self-motivation beliefs like self-efficacy. The performance control phase encompasses the processes that occur during learning, involving self-control strategies (e.g., attention focusing) and self-observation to monitor progress. The self-reflection phase follows performance and includes processes such as self-judgment (evaluating one's performance against a standard) and self-reaction (feelings of satisfaction and adaptive inferences that influence future efforts).

Multiple-choice questions (MCQs) provide students with a structured way to engage in performance control and self-reflection. They can self-assess their understanding, identify knowledge gaps, make causal attributions about their performance, and adapt their future learning strategies (Badali et al., 2023), thereby completing the SRL cycle. However, the manual creation of high-quality MCQs is a time-consuming and labor-intensive task for educators (Haladyna et al., 2002). Consequently, significant research efforts have been dedicated to developing techniques for automatic MCQ generation (Ch & Saha, 2020; Mulla & Gharpure, 2023). This research area has a history spanning more than two decades, with numerous MCQ generation systems developed across diverse languages and domains (Kurdi et al., 2020).

The integration of artificial intelligence (AI) into education is rapidly transforming teaching and learning methodologies, opening new possibilities for personalized learning, adaptive assessment, and data-driven teaching. One area of significant potential lies in the automated generation of educational resources, such as MCQs. Early MCQs generation systems relied on predefined templates and rule-based methods, limiting the scope and adaptability of the generated content. However, the rapid development of Natural Language Processing (NLP) models and the introduction of Large Language Models (LLMs) offer new possibilities to automatically generate these kinds of questions. The launch of ChatGPT in November 2022 and the rise of generative AI have opened concrete opportunities to effectively implement such systems in a real-world educational setting. Automated MCQs creation using generative AI offers a scalable alternative by producing questions tailored to course content and individual student progress with a simple prompt. However, concerns remain about the relevance and accuracy of AI-generated questions, as well as the students' perceptions of how such tools can support teachers in managing and improving the learning process.

To address these concerns, this study investigates the following research questions:

RQ1. *What is the level of relevance of MCQs generated by a generative AI-based system in a real-world computing education setting?*

RQ2. *How do students perceive the use of AI-generated MCQs in terms of usefulness, effectiveness, and overall learning experience?*

*RQ3. To what extent do LLMs fail to generate accurate multiple-choice questions in computing education courses and what types of errors can be identified?*

*RQ4. How does the design of AIQUIZ support the key phases of self-regulated learning?*

In the context of these research questions, we distinguish between relevance, usefulness, and accuracy. In RQ1, relevance refers to the perceived alignment of the MCQs with course content and learning objectives, quantified by student perception scores and the rate of 'Beyond the scope' flags in our error analysis. In RQ2, usefulness relates to the students' subjective experience of how the tool supports their overall learning process, including motivation, self-assessment, and study effectiveness. Finally, in RQ3, accuracy is strictly defined by the technical fidelity of the questions, measured through the overall student performance rate (the percentage of questions answered correctly) and the expert-validated error taxonomy used to classify reported flaws.

For this study, we have developed AIQUIZ, an open-source platform designed to create personalized MCQs using generative AI. This platform utilizes a prompt-based generation approach, which enables full control of the LLM parameters and prompts, relying solely on the LLM's inherent knowledge rather than fine-tuning, and allows us to access full information on user interactions, which is critical to analyze and understand the impact of these AI systems. To answer these research questions, we have applied this system in four distinct computing education courses at the Polytechnic University of Madrid (UPM), two on databases and two on web technologies. By exploring these aspects, this study seeks to contribute to a broader understanding of how generative AI can enhance educational practices.

The rest of the paper is structured as follows. Section 2 presents a review of related work on AI-generated MCQs. Section 3 describes the methodology of our study, detailing the

study design and data collection tools. Section 4 provides an overview of AIQUIZ, the platform used for generating AI-driven MCQs, including its architecture, functionalities, and prompts used. Section 5 presents the results and discussion, analyzing question performance, relevance and accuracy, student perceptions, and common errors. Finally, Section 6 summarizes the key findings, limitations and outlines directions for future research.

## 2. Related Work

An MCQ consists of the following components (Haladyna et al., 2002): 1) Stem: This is the portion of the question that provides the context or scenario upon which the question is based; 2) Key: The correct answer to the question; 3) Distractors: These are the incorrect options presented as plausible answers, designed to challenge and potentially mislead students. The creation of MCQs has traditionally been a manual and meticulous task. This is because it involves reviewing the source text, identifying relevant phrases that can be transformed into questions, selecting a word or expression as the correct answer, designing a question around that answer, and choosing distractors derived from the text or related context (Gamage et al., 2019; Haladyna et al., 2002).

However, with the development of NLP and advancements in AI, the automatic generation of MCQ has been greatly simplified, making it more accessible and efficient. In this context, a clear and promising opportunity arises to use LLMs for question generation, making them better aligned with educational objectives and the diverse needs of students. This approach has been successfully applied in different fields such as medicine (Kıyak et al., 2024; Mistry et al., 2024) or finance (Vu et al., 2024). In medical education, for instance, LLMs have been shown to produce high-quality, board-style questions for radiology that are comparable in quality to those created by human experts (Mistry et al., 2024). They have also demonstrated the ability to generate questions for pharmacotherapy exams that can effectively differentiate between high- and low-performing students (Kıyak et al., 2024). Similarly, in finance education, this approach is valued as a promising method to streamline the time and effort educators spend on creating assessments (Vu et al., 2024).

Several studies compare different LLMs as tools to generate MCQs, for example, a study conducted by Tran et al. (2023), compares GPT-3 and GPT-4 to create the answers given the question stem for an introductory low-level C programming course. The results were later evaluated for the accuracy of the generated answers, and GPT-4 outperformed GPT-3. Additionally, language models such as Llama 2, Mistral, and GPT-3.5 have been subjected to comparative analysis to generate these types of questions, with GPT-3.5 producing the most effective MCQs based on various established metrics (Biancini et al., 2024). Furthermore, technological advancements have facilitated the integration of LLMs such as GPT-4 with retrieval-augmented generation (RAG) (Lewis et al., 2020) to develop a generative AI framework specifically designed for automated MCQ creation (Hang et al., 2024), demonstrating its ability to align questions with specific learning objectives. In addition, Wang et al. (Wang et al., 2025) proposed a multi-agent workflow powered by LLMs, where critique agents iteratively refine questions to ensure alignment with pedagogical standards and readability, achieving an expert approval rate of 84.2%.

Further research (Cheung et al., 2023) compares the quality of MCQs produced by ChatGPT with those crafted by human experts in specialized domains like medical education. Although human-generated questions are more relevant, AI-generated MCQs

showed comparable overall quality, highlighting the potential of these models to match human performance in critical aspects such as difficulty and clarity. In all these studies, these models demonstrate the ability to generate high-quality and educationally effective MCQs on a large scale, provided that prompts are structured systematically and accurately.

Recent advancements have explored diverse AI-driven methodologies for generating MCQs in computing education, moving beyond simple template-filling to more sophisticated techniques. For instance, some research has utilized Conditional Generative Adversarial Networks (cGANs) to produce varied and contextually relevant questions across different proficiency levels, from beginner to expert (Shoaib et al., 2025). This approach leverages a generator and a discriminator in an adversarial process to create questions from a comprehensive dataset curated from textbooks and online resources. Other work has focused on specific components of the MCQ, such as the novel framework proposed by Kumar et al. (Kumar et al., 2024), which uses semantic and machine-learning techniques to automate the generation of question stems. In parallel, other studies have concentrated on harnessing the power of LLMs through refined prompting techniques. Song et al. (Song et al., 2024), for example, developed a system named EduCS which uses a "prompt chain" with GPT-3.5 to generate MCQs aligned with CS0 and CS1 curricula. Their findings underscore the potential of LLMs but also affirm that teacher verification remains crucial before questions are used by students. Further emphasizing the need for expert review, Grévisse et al. (Grévisse et al., 2024) conducted a docimological quality analysis of LLM-generated MCQs in computer science and medicine. Their study evaluated questions generated via zero-shot approaches against standard item-writing guidelines and identified common flaws such as ambiguous keys and implausible distractors, concluding that human oversight is essential for ensuring quality and instructional alignment. A distinct but related application is the generation of Questions about Learner's Code (QLCs), which are questions generated automatically about a student's own program (Lehtinen et al., 2023). One such tool for JavaScript generates multiple-choice questions targeting the structure and evaluation of a student's code to identify "unproductive success", instances where students produce functionally correct code but may lack a deep understanding of the underlying concepts.

A recent study by Doughty et al. (Doughty et al., 2023) explored the use of GPT-4 to automatically generate MCQs aligned with learning objectives in Python programming courses. Their approach involved a detailed generation pipeline, integrating Bloom's taxonomy and question types, and comparing 651 LLM-generated questions with 449 human-crafted ones. The study found that GPT-4 could produce questions with comparable quality and better alignment to learning objectives than human-authored items. While this represents a significant step forward in applying LLMs to programming education, the scope was limited to Python and the evaluation was conducted solely through expert annotation, without deployment in real classroom settings or direct feedback from students.

Overall, the literature highlights growing interest and progress in the automatic generation of MCQs using LLMs, particularly in domains such as medicine, finance, and programming education. However, key challenges remain unresolved, especially regarding the pedagogical relevance and factual accuracy of AI-generated questions, as well as the limited understanding of how students perceive and interact with such tools. Moreover, few studies have systematically analyzed student-reported feedback on these questions at scale. Our study addresses this gap by examining student-flagged errors in

programming and database courses, offering a practical perspective on the use of LLM-generated MCQs in real classroom settings.

## 3. Methodology

This section outlines the methodology employed in this study. The study was conducted in a real-world educational setting at the Polytechnic University of Madrid, focusing on the use of AI-generated MCQs in computing courses in higher education.

### 3.1 Study Design

The primary objectives of this study were to answer our four research questions. Specifically, we aimed to: (1) assess the relevance (i.e., content alignment with the course) of the AI-generated MCQs (RQ1); (2) evaluate student perceptions of the platform's usefulness as a learning aid and their overall experience (RQ2); (3) analyze the specific types of errors the LLM produces (RQ3); and (4) examine how the design of AIQUIZ supports the key phases of self-regulated learning (RQ4). For the evaluation of relevance (RQ1), we relied solely on the LLM's baseline knowledge without additional pre-training. No course-specific materials were used, and the topics were not tailored or mapped to the actual course content. The study was conducted over two academic years (2023/24 and 2024/25) and involved students enrolled in four distinct undergraduate computing engineering courses. "Network Computation" and "Web Engineering" courses of the Bachelor´s Degree of Telecommunications Engineering, "Databases" course of the Bachelor´s Degree of Biomedical Engineering, and "Non-Relational and Distributed Databases" course of the Bachelor´s Degree of Science in Data Engineering and Systems.

The courses were grouped into two main subjects: 1) Web Technologies, where students learn web application development, both on the client side (HTML, CSS, JavaScript, React, React Native) and the server side (Node.js, Express) and 2) Databases, where students learn the characteristics of non-relational databases, families of non-relational databases, and study the MongoDB database in depth.

A total of 593 students were enrolled across the four courses (415 in Web Technologies and 178 in Databases). Of these, 325 (54.8%) used AIQUIZ at least once and constitute the effective user base analysed in this study; the remaining 268 enrolled students did not use the platform. Throughout this paper we distinguish between enrolled students (593), active users of the platform (325), and survey respondents (47). Unlike small-scale pilot studies, this large-scale deployment ensures high ecological validity, providing a robust empirical basis (over 38,000 interactions) to analyze the system's impact in a genuine curricular environment. The courses were conducted face-to-face over a four-month period, with an average duration of 130 working hours, equivalent to 4.5 European Credit Transfer and Accumulation System (ECTS) credits. These courses were selected due to their alignment with the capabilities of LLMs in generating domain-specific questions.

At the beginning of each course, students were introduced to AIQUIZ, a web-based platform designed for the automatic generation of MCQs using generative AI, and were encouraged to use it as a supplementary tool for self-assessment outside of class hours. To familiarize them with the tool, a brief demonstration was provided in class, showcasing its intuitive interface and core functionalities. A key ethical consideration in this study was the potential for the AI to generate incorrect questions or answers, which could mislead students. To mitigate this risk, the platform was explicitly positioned as a

voluntary self-assessment tool, and its use had no impact on student grades, reducing the stakes associated with encountering erroneous content. Students were also informed that they were participating in a research study evaluating a new AI tool, which encouraged a critical perspective.

AIQUIZ provides full control over the configuration of LLMs, allowing precise adjustments. For this study, OpenAI's GPT-4o Mini was used as the underlying LLM. The model was configured with a temperature setting of 1.0 (on a scale from 0 to 2), promoting a balance between creativity and coherence in question generation. Additional parameters included a frequency penalty of 0 and a presence penalty of 0, ensuring that the model did not overly favor or avoid specific terms. The maximum token limit was set to 2048 to allow for comprehensive question formulation and explanations, while a single response was generated per request (n = 1). This configuration aimed to optimize both the adaptability and accuracy of the generated MCQs.

### 3.2 Data Collection Tools

The primary data collection tool used in this study was AIQUIZ. The platform logged all interactions, including the number of questions generated, student responses, and the accuracy of those responses. Additionally, AIQUIZ enabled students to report any errors or issues with the generated questions, providing valuable feedback for further analysis. The platform's architecture, functionality, and user interface are described in detail in the subsequent section.

To evaluate student perceptions of AIQUIZ, a voluntary survey was administered. The survey consisted of 13 questions using a 5-point Likert scale to assess aspects such as usability, perceived usefulness, accuracy of questions, and overall satisfaction. The detailed survey results are presented in the results section (Figure 3). Given that the survey questions and results are displayed in the results section, they are not repeated here. The survey instrument was subjected to an exploratory psychometric evaluation. First, the internal structure of the AIQUIZ evaluation questionnaire was examined using exploratory factor analysis (EFA). Prior to factor extraction, the suitability of the data for factor analysis was assessed. Bartlett's test of sphericity was significant ($p < .05$), indicating that the correlation matrix was not an identity matrix, and the Kaiser–Meyer–Olkin measure of sampling adequacy showed a value of 0.83, suggesting good factorability of the data. The dataset was randomly split into a model-building subsample and a holdout subsample. The number of factors to retain was determined using parallel analysis, which indicated a single-factor solution. Based on this result, EFA models with increasing numbers of factors were estimated using maximum likelihood with robust standard errors (MLR) and geomin rotation. However, the one-factor solution was retained as the most parsimonious and empirically supported representation of the data. The retained factor was interpreted as representing overall evaluation of AIQUIZ. Because the questionnaire included both positively and negatively worded items, items referring to errors or mistakes in the tool (Q4, Q6, Q7, Q8) were reverse-coded so that higher scores consistently reflected more positive evaluations. To ensure the rigorous empirical validity of the survey instrument, a psychometric evaluation was conducted. Internal consistency of the resulting unidimensional scale was assessed using Cronbach's alpha, which showed good internal consistency (alpha = 0.79).

Of the 325 students who actively used the platform, 47 completed the survey, representing a response rate of 14.5% relative to active users (7.9% relative to the 593 enrolled students). This limited response rate should be considered when interpreting survey

results, as respondents may disproportionately represent students with higher engagement or more positive experiences with the platform (self-selection bias).

Finally, five professors manually reviewed the reported MCQs to validate and classify the types of errors that the students had encountered. These professors categorized the reported questions based on specific issues such as confusing wording, poorly formulated options, or incorrect answers. This process allowed for a detailed analysis of the types of errors present in the AI-generated questions and provided insights into areas for improvement.

## 4 AIQUIZ

### 4.1 Overview

AIQUIZ is an open-source web-based platform designed for the automatic generation of MCQs across various computing education courses. The platform provides comprehensive control over the configuration of large language models (LLMs), including prompt customization and parameter tuning, ensuring a consistent and standardized user experience for students. This approach offers several advantages over direct interaction with AI systems like ChatGPT. First, AIQUIZ enables the creation of controlled experimental settings where all students use the same prompt format. Additionally, the platform supports the generation of MCQs tailored to students' progress, facilitating personalized learning experiences. By centralizing the use of a single LLM model for all students, AIQUIZ ensures equitable access while eliminating dependency on individual subscriptions or service fees. Finally, the platform records user interactions and relevant data exchanged with the AI API, providing valuable insights for further analysis.

### 4.2 User Interface and Functionalities

AIQUIZ's main functionality is the generation of adaptive MCQs. Four screen captures can be seen in Figure 1. On the first screen, the student can choose the corresponding subject. Then, a screen with 4 options is presented, where the student has to choose the topic and subtopic, the difficulty of the questions, and the number of questions to generate, and request them by clicking on the “Create test” button. Then the student obtains a new screen with the generated questions. In each question, the student can choose the correct answer and click on the “Answer” button or click the “Report incorrect question” button to report any problem related to the question. Once an answer is submitted, AIQUIZ provides immediate feedback by displaying an additional box containing a detailed explanation of the correct answer to the MCQ. AIQUIZ will also show a final feedback screen including the final grade obtained and a message encouraging to retry it when finishing all the questions. There is an extra welcome screen the first time the student accesses AIQUIZ, where he/she has to introduce a valid email that will be used to track progress and save all interactions.

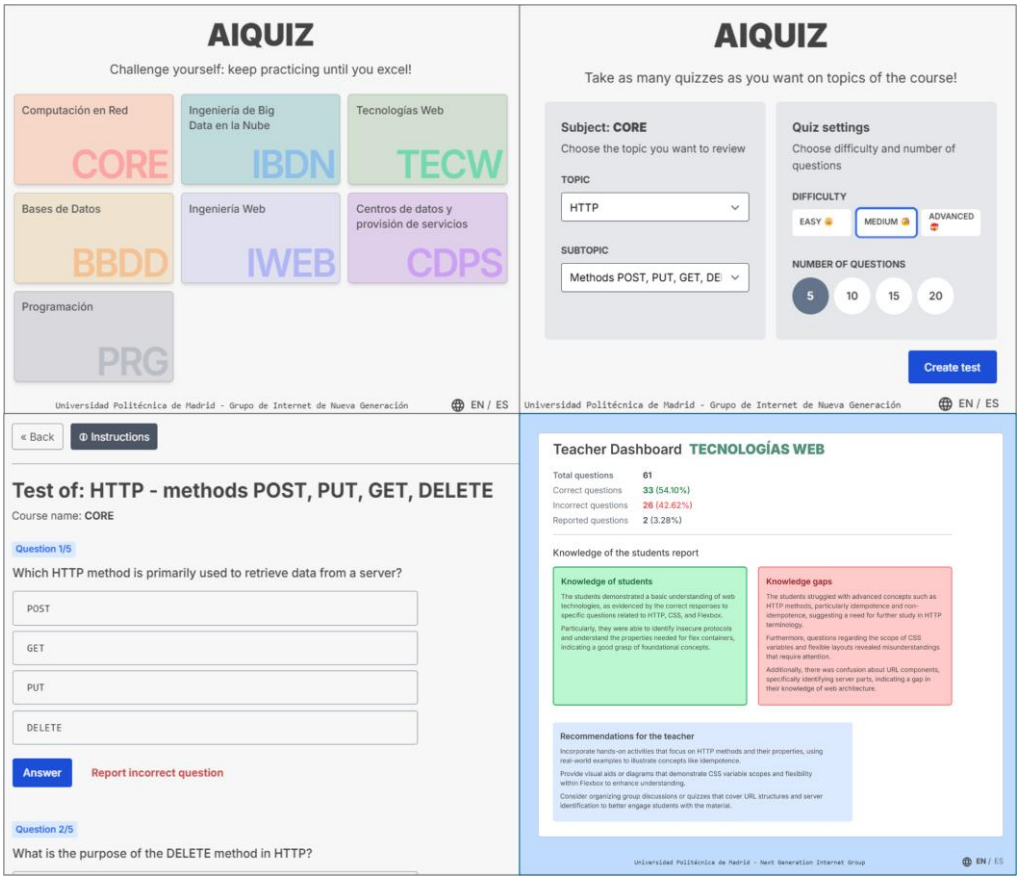
The fourth screen shows a subject dashboard only accessible to the teacher. This dashboard is also generated using artificial intelligence with a special prompt. It shows recommendations and knowledge gaps that are detected taking into account the student answers to the MCQs.

It is important to emphasize that the platform operates independently of teacher-created materials, relying instead on predefined generic topics. This option is only suitable for

subjects such as programming and databases, domains that are well-supported by LLMs. This enables the automatic generation of MCQs through appropriately crafted prompts without requiring additional subject-specific customization.

**Figure 1**.

*Four screen captures of AIQUIZ*




### 4.3 Prompts

The prompt that the AIQUIZ server uses to generate the MCQs is "I am a higher education student enrolled in the subject {subject}. Generate {number} multiple choice questions on the topic {topic} in the knowledge domain {domain}. I have previously answered {number2} questions, which are provided here with their corresponding answers: {questions}. Use my past responses to create new questions that help deepen my understanding of the topic. The questions should be at a {difficulty} level of difficulty. Ensure that each correct answer corresponds accurately to its respective question. {rest}.`. The AIQUIZ backend customizes this prompt by replacing the placeholders with specific values:

- {subject}: Replaced with the subject name selected by the student on the initial screen (Figure 1).
- {number}, {topic}, {domain}, {difficulty}: Replaced with form parameters provided by the student on the second screen.

- {number2}: Replaced with the number of previously answered questions, while {questions} includes the corresponding questions, answers, and whether the student's responses were correct, extracted from the MongoDB database.
- {rest}: Specifies the desired response format in JSON, including specific parameters required for correctly rendering the result.

The prompt that is used to generate the dashboard is "From a question bank for the subject {subject} of an engineering degree at the Polytechnic University of Madrid, {number} questions have been answered. The questions are about {topics} topics. There have been {number2} questions answered correctly. Students have answered {number3} questions incorrectly, which are as follows: {questionswrong}. Make a small report in markdown format with a paragraph indicating the "Knowledge of the students" and another one the "Knowledge gaps", i.e. the topics where they fail the most. Add a third paragraph with the "Recommendations for the teacher" of the subject with tips and ideas to help students improve their knowledge." As before, this prompt is customized by the AIQUIZ backend, by substituting the following texts:

- {subject} and {topics} by the subject name and its topics.
- {number} by the number of questions answered.
- {number2} and {number3} by the number of questions answered correctly and incorrectly.
- {questionswrong} by the questions answered incorrectly on the topic. If the number of such questions exceeds 20, a random subset of 20 questions is selected from the incorrectly answered questions on the topic.

All this data is obtained from the database and config files. An important consideration is that student data is never shared with the AI provider during the prompting process. All interactions are managed through the AIQUIZ server, which acts as the sole client from the perspective of the LLM provider.

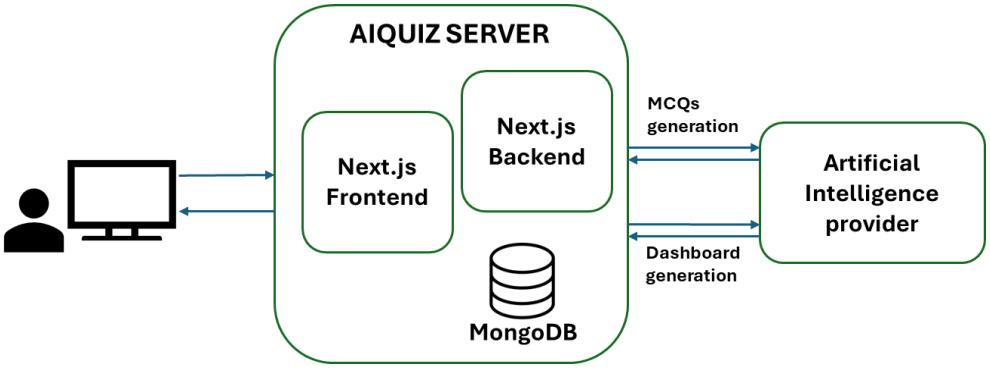
### 4.4 Architecture

The AIQUIZ platform is built using Next.js, an open-source web development framework that facilitates the development of full-stack web applications by combining React for the frontend and Node.js for the backend. AIQUIZ is an open-source project available on GitHub at https://github.com/CyberAula/aiquiz. Its system architecture, illustrated in Figure 2, integrates several components to support its functionalities.

In addition to its frontend and backend layers, AIQUIZ employs a MongoDB database to store user interactions and data exchanged with the AI provider. This persistent storage enables detailed logging and facilitates data-driven analysis. The backend serves as the intermediary between the platform and the AI provider, handling requests for generating MCQs and producing a teacher dashboard.

**Figure 2.**

*AIQUIZ server architecture*




Unlike other MCQ platforms, AIQUIZ´s design focuses on the student's learning process rather than just question generation. By considering previous attempts, AIQUIZ tailors questions to reinforce understanding, making learning more personalized and effective. Additionally, the platform empowers educators with insightful reports, helping them identify common misconceptions and areas for improvement. AIQUIZ also integrates a mechanism that allows students to report errors, so they can be reviewed by teachers. This holistic approach positions AIQUIZ not just as a question generator but as a tool designed to support self-regulated learning.

### 4.5 AIQUIZ and Support for Self-Regulated Learning

The design of the AIQUIZ platform intentionally incorporates features that align with the three phases of Zimmerman's SRL model (Zimmerman, 2000) to create a comprehensive learning tool.

The platform promotes the **forethought phase** by prompting students to engage in goal setting and strategic planning. Before generating a quiz, students must choose the subject, topic, difficulty level, and number of questions (see Figure 1). This process encourages them to actively consider their learning objectives and tailor the practice session to their perceived needs.

During the quiz, students engage in the **performance control phase**. The act of answering questions requires attention focusing and the application of task strategies to recall and apply knowledge. The platform's adaptive question generation, which uses past responses to create new questions, acts as a strategy to help students deepen their understanding. Furthermore, by tracking their own performance in real-time, students engage in self-observation, a key element of performance control.

Finally, AIQUIZ is designed to facilitate the **self-reflection phase** through immediate feedback. After each question, the platform indicates whether the answer was correct and provides an explanation, allowing for immediate self-evaluation. The "Report incorrect question" feature is also a tool for self-reflection, as it encourages students to critically assess the validity of the question and its answer, a form of adaptive inference.

## 5 Results And Discussion

This section presents the study's findings organized by the four research questions defined in the Introduction. First, we analyze the relevance and accuracy of the generated

questions (RQ1 & RQ3), followed by an examination of student perceptions and acceptance (RQ2). Finally, we synthesize these findings to discuss how the platform supports the phases of Self-Regulated Learning (RQ4). Although the AIQUIZ platform is adaptable to various educational domains through the use of configuration files, the present study is firmly situated within the context of computing education. All prompts used during the evaluation were specifically designed for computing courses, drawing upon domain-specific terminology and concepts relevant to the courses. For example, prompts referenced programming languages (e.g., JavaScript, Python), database operations and syntax (e.g., MongoDB queries, NoSQL data models), and web development frameworks (e.g., React, Node.js) aligned with the course content. Importantly, AIQUIZ did not rely on teacher-created materials or custom datasets, instead, it leveraged the inherent knowledge of LLMs (using a prompt-based generation approach), which are particularly effective in domains like computing, where the underlying concepts, terminology, and patterns are extensively covered in the training data of large language models. This domain-specific design ensures that the results and findings are directly applicable to computing education.

### 5.1 Question Relevance and Accuracy

To date, a total of 38,752 questions have been generated by the 325 active users (54.8% of the 593 enrolled students): 26,613 in the Web Technologies courses (of the 415 enrolled students, 211 have used it [50.84%]) and 12,139 in the Databases courses (of the 178 enrolled students, 114 have used the tool at least once [64.04%]). This corresponds to an average of approximately 119 questions per active user (≈126 in Web Technologies and ≈107 in Databases). With the gathered data, we try to answer the first research question, "*What is the level of relevance of MCQs generated by a generative AI-based system in a real-world educational setting?".*

Overall, students correctly answered 70.79% of the generated questions, with notable differences between courses: Databases had a higher performance rate (79.45%) compared to Web Technologies (66.84%). It is important to note that student performance rate should not be conflated with question quality. These are distinct dimensions: performance reflects a combination of prior knowledge, item difficulty, and student familiarity with the topic, rather than serving as a direct indicator of technical question quality or construct validity. The 70.79% rate suggests that questions were at an appropriate difficulty level and topically relevant, but it does not, by itself, confirm that questions were well-formulated. This difference between courses may be due to the fact that the courses belong to different programs and levels. Additionally, the study did not control for the intrinsic difficulty or scope of the different course curricula when comparing performance across Web Technologies and Databases, as the analysis focused on the observed performance within each distinct course setting. Web Technologies courses correspond to more advanced levels and cover broader concepts. Furthermore, in Databases, course teaching staff recommended that the students choose medium-difficulty questions, whereas in Web Technologies, this recommendation was not made, and they selected high-difficulty questions in 32.22% of the cases and medium-difficulty questions in 64.55%. Another important factor that may explain the difference in performance between Database and Web Technologies courses is the student profile. The Web Technologies courses are offered in the Bachelor's Degree in Telecommunications Engineering, which generally has a lower entry score compared to the Bachelor's Degrees in Biomedical Engineering and Data Engineering and Systems, from which the Database courses are drawn. Consequently, students entering these latter degrees tend to have

stronger academic preparation on average which may have led to a better performance in answering MCQ tests.

In terms of engagement, the distribution of use among the 325 active users was highly skewed: they answered a median of 65 questions, well below the mean of approximately 119, with a single student accumulating 1,545 questions alone. The median is therefore reported as a more representative measure of central tendency. This highly active student was not excluded from the study, as the large number of answered questions contributed to the model evaluation. Table 2 collects the most active subtopics in each subject, with the performance rate and the difficulty chosen.

To evaluate the tool's performance, students had the option to report errors in the generated questions. This feature provided valuable insights into the accuracy of AIQUIZ and the reliability of AI-generated MCQs. Overall, only 327 questions (0.85%) were flagged as potentially incorrect. Though this low number reflects a positive perception of the tool's reliability from the user's standpoint, it should be interpreted with caution, as students may not always be able to accurately identify errors. Notably, Databases, the subject with the highest correct answer rate (79.45%), also had the highest proportion of reported errors relative to the number of generated questions (1.19%), compared to 0.69% in Web Technologies.

Based on the number of questions generated, students showed the highest engagement in the topics "Introduction to NoSQL" and "MongoDB Shell" within the Databases courses, and "URLs" and "Creation and Usage of Components in React" in Web Technologies. An analysis of student performance revealed the topics that posed the most significant challenges. In the Databases courses, students achieved the lowest performance on questions related to the "Unwind Operator" and "Update Operator," with correct answer rates of only 64.11% and 69.68%, respectively.

This suggests students struggled to master MongoDB's aggregation pipelines and update operator semantics. Similarly, in Web Technologies, students performed most poorly on "CSS Grid" and "Node file system," answering only 47.46% and 49.40% of questions correctly, which indicates difficulties with core front-end concepts like CSS rule precedence and runtime behavior.

Some examples of questions students failed are shown in Table 1. Table 3 summarizes the subtopics in which students faced more difficulties in each subject. With this data, teachers can access in a rapid and easy way the topics that cause more problems to the students.

**Table 1.**

*Examples of questions students failed*

1) [Web Technologies, CSS – Grid] Which of the following CSS properties is used to set the number of columns in a grid?
    a. :grid-rows
    b. :grid-columns -> Student answer
    c. :grid-template-columns -> Correct answer
    d. :grid-gap
2) [Web Technologies, Node – File System] What method is used to delete a file in Node.js asynchronously?
    a. fs.deleteFile()-> Student answer
    b. fs.remove()

c. fs.unlink() -> Correct answer
d. fs.erase()

3) [Databases, MongoDB Aggregation Framework, Unwind Operation] What projection operator is used to split a field into multiple fields in the output documents?
   a. $split -> Correct answer
   b. $divide -> Student answer
   c. $break
   d. $explode
4) [Databases, MongoDB Shell – Update Operator] When performing an update in MongoDB, which parameter is used to specify that the operation should create a new document if no document matching the filter is found?
   a. upsert -> Correct answer
   b. insert -> Student answer
   c. createIfNotFound
   d. addIfMissing

These patterns are consistent with prior research showing that students in introductory programming often struggle with syntax, conceptual understanding, and strategic application of knowledge (Qian & Lehman, 2017). Such difficulties stem not only from the technical complexity of programming but also from students' prior misconceptions, fragile mental models, and beliefs about programming aptitude and self-efficacy, which significantly affect their persistence and performance in computing courses (Tek et al., 2018). AIQUIZ surfaces these domain-specific issues and proves useful not only as an assessment tool but also as a lens for diagnosing learning barriers unique to computing education. In doing so, it opens up promising directions for future research focused exclusively on identifying and analyzing conceptual difficulties in computer science education at scale using data from platforms like AIQUIZ.

The high overall performance (70.79%) suggests that AIQUIZ may be useful for supporting the performance control and self-reflection phases of SRL. Receiving predominantly correct feedback can enhance students' self-efficacy, a core component of the forethought phase. Furthermore, the low error reporting rate (0.85%) is consistent with the platform's potential value for student self-evaluation, a cornerstone of the self-reflection process, as it builds trust in the feedback provided.

**Table 2.**

Subtopics with more questions generated in each subject.

| Subject | Topic | Subtopic | Total Questions | Performance | Difficulty % | | |
|---|---|---|---|---|---|---|---|
| | | | | | Easy | Medium | High |
| Web Technologies | HTTP | URLs | 1,207 (4.54%) | 69.10% | 8.53% | 62.88 % | 28.58 % |
| | React | Creation and usage of components | 1,072 (4.03%) | 64.27% | 16.32 % | 66.42 % | 17.26 % |
| | React | React Native | 964 (3.62%) | 71.89% | 4.25% | 76.97 % | 18.78 % |
| | HTTP | HTTP Petition Format | 948 (3.56%) | 76.90% | 5.06% | 60.34 % | 34.60 % |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | React | Props and State | 750 (2.81%) | 63.20% | 4.67% | 73.6% | 21.73 % |
| Databases | Introduction to NoSQL | Data lifecycle management | 849 (6.99%) | 76.33% | 0% | 100% | 0% |
| | Introduction to NoSQL | The 5Vs of Big Data | 735 (6.05%) | 89.39% | 0% | 100% | 0% |
| | Introduction to NoSQL | Data value pyramid | 725 (5.97%) | 70.48% | 0% | 100% | 0% |
| | MongoDB Shell | Basic commands | 714 (5.88%) | 79.55% | 0% | 100% | 0% |
| | Introduction to NoSQL | CAP theorem | 704 (5.80%) | 76.99% | 0% | 100% | 0% |

**Table 3.**

Subtopics with more errors in each subject.

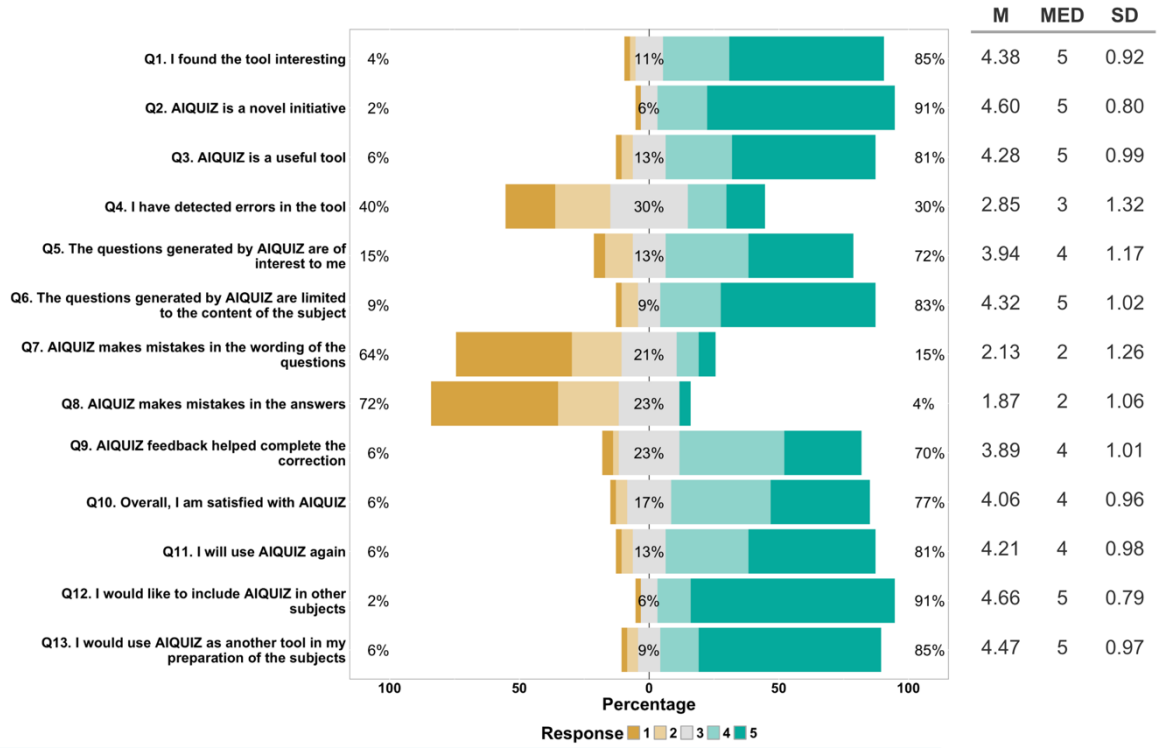
| Subject | Topic | Subtopic | Total Questions | Performance | Difficulty % | | |
|---|---|---|---|---|---|---|---|
| | | | | | Easy | Medium | High |
| Web Technologies | CSS | Grid | 59 (0.22%) | 47.46% | 0% | 83.05% | 16.95% |
| | CSS | Positioning | 89 (0.33%) | 49.44% | 0% | 77.53% | 22.47% |
| | Node | File System | 83 (0.31%) | 49.40% | 0% | 71.08% | 28.92% |
| | Client JavaScript | Document object and windows | 185 (0.70%) | 49.73% | 0% | 50.27% | 49.73% |
| | Client JavaScript | DOM and DOM Manipulation | 227 (0.85%) | 51.10% | 2.20% | 60.35% | 37.44% |
| Databases | MongoDB Aggregation Framework | Unwind Operator | 482 (3.97%) | 64.11% | 0% | 100% | 0% |
| | MongoDB Shell | Update Operations | 442 (3.64%) | 69.68% | 0% | 100% | 0% |
| | Introduction to NoSQL | Data value pyramid | 725 (5.97%) | 70.48% | 0% | 100% | 0% |
| | Schema Design in NoSQL and MongoDB | No relational data model or NoSQL | 596 (4.91%) | 74.83% | 0% | 100% | 0% |
| | MongoDB Shell | Delete operations | 354 (2.92%) | 76.27% | 0% | 100% | 0% |

### 5.2 Student Perceptions and Acceptance

A voluntary survey was conducted among the students of both subjects to gather their perceptions on AIQUIZ. Figure 3 presents the results of the survey, which consisted of 13 questions using a 5-point Likert scale, where 1 = Strongly Disagree, 2 = Disagree, 3 = Neither Agree nor Disagree, 4 = Agree, and 5 = Strongly Agree. The survey was completed by 47 students from both subjects. The results were aggregated for both subjects due to their significant similarities and the limited number of respondents.

**Figure. 3.**

*Results of the student survey (N=47)*




With this data, we answer the second research question "*How do students perceive the use of AI-generated MCQs in terms of usefulness, effectiveness, and overall learning experience?". Relying on the psychometrically validated instrument described in Section 3 (Cronbach's alpha = 0.79), the analysis provides statistically reliable evidence of the tool's impact.*

The survey results reveal a strong approval of the AIQUIZ platform among students (Q1: mean = 4.38, SD = 0.92, MED = 5), emphasizing its engagement (Q11: mean = 4.21, SD = 0.98, MED = 4) and satisfaction (Q10: mean = 4.06, SD = 0.96, MED = 4). Students expressed that AIQUIZ was a valuable resource, considering it both a novel initiative (Q2: mean = 4.60, SD = 0.80, MED = 5) and a useful tool (Q3: mean = 4.28, SD = 0.99, MED = 5). Furthermore, they reported detecting relatively few errors (Q4: mean = 2.85, SD = 1.32, MED = 3).

Regarding the questions generated, students stated that they found them interesting (Q5: mean = 3.94, SD = 1.17, MED = 4) and aligned with the subject topics (Q6: mean = 4.32, SD = 1.02, MED = 5). Students identified areas for improvement, such as infrequent errors in the answers (Q8: mean = 1.87, SD = 1.06, MED = 2) or occasional issues with the wording of the questions (Q7: mean = 2.13, SD = 1.26, MED = 2). However, these

issues were seen as relatively minor and did not significantly detract from the platform's overall utility. Notably, there was considerable appreciation for the feedback mechanism, which students found helpful in completing and clarifying their corrections (Q9: mean = 3.89, SD = 1.01, MED = 4).

The survey also indicates a strong willingness among students to continue using the platform (Q11: mean = 4.21, SD = 0.98, MED = 4), including incorporating it as a new tool for studying their courses (Q13: mean = 4.47, SD = 0.97, MED = 5). Additionally, they expressed a desire to adopt the platform in other subjects (Q12: mean = 4.66, SD = 0.79, MED = 5). These findings are consistent with recent research on student beliefs about generative AI in education (Cabellos et al., 2026), which reports positive attitudes and constructivist-oriented use among students.

Engagement may be partially explained by the prompt design, which includes correct and incorrect responses from previous tests, providing the LLM with context to generate varied questions rather than repeating similar formulations. Additionally, students can select different difficulty levels, which may also produce different questions. These design choices likely contributed to the highly positive perception reported in survey questions Q5 and Q6. While the primary focus of this study was on the reliability and overall usability of AI-generated MCQs, it is important to note that perceived usability can also depend on the diversity of generated questions. Future work could explore dynamic adjustment of generation parameters, such as temperature and penalties, and evaluate how these adjustments affect both question diversity and pedagogical usefulness.

Overall, the survey results, supported by the instrument's strong construct validity, suggest that students perceive AIQUIZ as a valuable tool that enhances their learning experience. These findings also provide empirical support for RQ4, providing perception-level support for the conceptual claims regarding SRL. As students' positive perceptions are consistent with conditions associated with key phases of SRL. The strong student approval and high engagement levels (Q1, Q11, Q13) are consistent with the motivational conditions associated with the forethought phase of SRL, as positive perceptions of the tool's usefulness and novelty likely enhance students' intrinsic motivation and self-efficacy beliefs, making them more willing to engage in self-regulated learning activities. The appreciation for the feedback mechanism (Q9) suggests the platform may play a role in supporting the self-reflection phase, as students found it helpful for clarifying corrections and reinforcing their understanding. Finally, students expressed a strong willingness to continue using AIQUIZ and extend its application to other subjects, highlighting its potential as a scalable educational resource aligned with the cyclical nature of SRL.

### 5.3 Error Analysis of AI-Generated MCQs

In this study, a total of 327 questions (0.85% of all generated MCQs) were reported by students as potentially incorrect. To address the third research question, *"To what extent do LLMs fail to generate accurate multiple-choice questions in computing education courses, and what types of errors can be identified?"* five professors manually reviewed the reported questions, distributing them randomly among themselves. Each question was categorized based on predefined error types, divided into (1) issues related to the formulation of the question and options (including confusing wording, poorly formulated options, repeated options, multiple correct answers, missing correct answers); (2) issues with the correct response according to the LLM (incorrectly marked answers, erroneous explanations); and (3) questions beyond the course scope. To validate the robustness and

empirical reliability of the human evaluation process, we conducted a rigorous inter-rater agreement analysis. We randomly selected 65 questions per subject, each of which was evaluated by two experts. We then computed Cohen's kappa separately (Landis & Koch, 1977) for each of the four categories: "Everything correct" and each of the three possible issue types, with more than a $\kappa = 0.6$ for moderate agreement between the raters. In all evaluations, a Cohen's kappa value above 0.61 was obtained, indicating an acceptable level of agreement among the experts (Graham et al., 2012).

Of the 327 reported questions, 98 (29.97%) were determined by professors to be correct, meaning that nearly a third of student-reported errors were false positives. The distribution of these cases was similar across subjects, with 31.25% of reported questions in Databases (inter-rater agreement of $\kappa = 0.87$) and 28.96% in Web Technologies ($\kappa = 0.74$) ultimately deemed valid. Table 4 summarizes the distribution of confirmed errors among the remaining 229 incorrect questions.

**Table 4.**

*Issues of reported questions that the teachers have confirmed as incorrect.*

| Issue | Web Technology (Inter-rater agreement [κ]) | Databases (Inter-rater agreement [κ]) | Total |
|---|---|---|---|
| Any issue | 130 ($\kappa = 0.74$) | 99 ($\kappa = 0.87$) | 229 |
| Issues with question formulation | 100 (79.92%) ($\kappa = 0.66$) | 68 (68.69%) ($\kappa = 0.63$) | 168 (73.36%) |
| Issues with the correct response according to the LLM | 38 (29.23%) ($\kappa = 0.61$) | 43 (43.43%) ($\kappa = 0.66$) | 81 (35.37%) |
| Out of subject | 5 (3.85%) ($\kappa = 1.00$) | 4 (4.04%) ($\kappa = 1.00$) | 9 (3.93%) |

These findings align with established research on the challenges of writing effective MCQs, which identifies similar issues as common pitfalls for human authors (Haladyna et al., 2002; Purchase et al., 2010). It is important to note that the system not only fails to formulate questions, but also fails to respond to the questions it generates with a rate of 35.37%, being especially relevant in Databases (43.43%). However, errors in the explanation of the answer are not as frequent (18.34%), which suggests that the tool mainly fails in marking the correct answer, but if the explanation is read, it is accurate. This mitigates the impact of the failures as the explanation can help students to detect the mistake. The following example shows an incorrectly marked answer but a correct explanation:

- **Question**: What attribute is used to create a link to an email address in HTML?
- **Options**: "url", "address", "mail" (answer chosen by the LLM), "href" (correct answer)
- **Explanation:** The "href" attribute is used to establish a link to an email address in HTML using the <a> tag.

And this example shows an incorrectly marked answer and explanation:

- **Question**: What is the command in MongoDB Shell to display all existing databases?
- **Options**: “show databases()” (answer chosen by the LLM), “show dbs” (correct answer), “list dbs”, “get dbs”
- **Explanation:** The correct way to display all existing databases in MongoDB Shell is by using the show databases() command.

The most common issue is that the tool makes errors during the question generation process with a rate of 73.36%, mainly affecting Web Technologies with 79.92%. Among the causes reported by the experts is that the options provided do not present any challenge to the students, as seen in the following example:

- **Question**: "What is the spread operator in JavaScript, and what is it used for?"
- **Options**: "Promises", "Prototypes", "Try/catch", "Spread Operator" (correct answer/answer chosen by the LLM)

While this issue impacts the user experience, it is not critical because it does not confuse the students. In other cases, no correct option exists:

- **Question:** Which of the following characteristics is not part of the ACID properties of databases?
- **Options:** “Atomicity”**, “**Consistency”**, “**Isolation”**, “**Durability” (answer chosen by the LLM)

In other cases, multiple correct options are present:

- **Question:** How can I compare two values in JavaScript?
- **Options:** “val1 = val2”, “val1 == val2” (answer chosen by the LLM / correct answer), “val1 === val2” (also correct answer), “val1 != val2”

A similar issue, but rarely present (5.68%), happens when the tool repeats options in their formulation. For example, in the question "What is the tag used to create a comment in an HTML document?" the option "<!-->" is repeated twice. In this category we have also detected questions are formulated in a way that is difficult for students to understand. For example, "What is the most common way to declare a variable in JavaScript?" with the options "var, let, const, variable" creates confusion by asking for the "most common way" instead of the recommended way. Finally, almost no questions were found to be out of the scope of the subject (3.93%), indicating that the tool aligns well with the course content. An example of an out-of-scope question is the question “When is the ‘componentDidUpdate’ method called?” because this method is not explained in the Web Technologies course.

The error analysis in our study reveals both similarities and key differences when compared to the quality issues found in crowd-sourced MCQ repositories, such as the PeerWise platform analyzed by Purchase et al (Purchase et al., 2010). While both AI-generated and student-generated questions suffer from similar high-level flaws, the nature and distribution of these errors differ. Specifically, our findings align with those from the PeerWise study on several key quality issues. First is the issue of incorrectly marked answers. Both AIQUIZ and the student-authored questions in PeerWise occasionally designated the wrong option as the correct answer. The PeerWise study also identified this as an issue, noting that middle-quartile students were more likely to make this error, possibly by over-extending their abilities on complex topics. Another issue is confusing or unclear wording. We identified "confusing wording" in 16.16% of reported errors.

Similarly, the PeerWise study noted that while most questions were clear, higher-performing students sometimes struggled with clarity when creating more complex, multi-topic questions. Our analysis found that "poorly formulated options" was a frequent error. This mirrors the PeerWise study finding that creating feasible distractors is a difficult task, with only higher-performing students consistently succeeding. The PeerWise study highlighted that providing a good explanation was a significant challenge, even for the highest-quartile students. While our study found that the AI's explanation was often correct even when the marked answer was wrong, errors in explanations still occurred, indicating a common point of failure for both human and AI generators.

The fact that students reported 327 questions indicates they were not passive recipients but were actively engaged in self-monitoring and self-evaluation. Although nearly 30% of these reports were false positives, the act of questioning the material is itself a valuable reflective practice. Issues with the correct response according to the LLM (35.37%), poses a risk to accurate self-evaluation. However, since the accompanying explanation was often correct, students who read it could still make accurate causal attributions about their knowledge, thus completing the reflective cycle. It is important to note that among the 38,752 questions generated, there may be additional incorrect questions that were not reported by students (false negatives), meaning the actual error rate may be slightly higher. These findings suggest that 'full automation' in educational assessment is currently a misnomer. Instead, the most effective paradigm is a 'Human-in-the-Loop' workflow where AI handles volume and variety, while educators provide critical oversight. Future work should explore additional validation methods, such as automatic detection mechanisms (maybe using another LLM) or periodic expert reviews, to further ensure question quality.

### 5.4 Implications for Research and Teaching Practice

Our findings, particularly the discrepancy between high student acceptance and the persistence of specific error types, present several important implications for both future research and teaching practices. For the research community, this study offers a replicable mixed-methods framework for evaluating AI-driven educational tools in real-world settings, combining large-scale usage data with student perceptions and a detailed error taxonomy.

AIQUIZ is designed to support conditions associated with key self-regulated learning (SRL) strategies across the three phases of Zimmerman's cyclical model. In the forethought phase, the strong acceptance and motivation reported by students suggest that the platform may support the conditions for the planning and goal-setting processes that precede learning. During performance control, the high engagement levels are consistent with active self-monitoring of progress. Finally, the error-reporting mechanism provides direct support for the self-reflection phase: students who flagged potentially incorrect questions were actively engaging in self-evaluation, a core metacognitive process regardless of whether the reported item was ultimately confirmed as incorrect.

A key limitation of this study is that it does not evaluate teacher usage of the platform or their perception of its usefulness in the courses. In this regard, teachers tend to have a more critical view than students. As noted, students may fail to identify poorly formulated questions, which can lead to slightly lower performance of the platform. Moreover, teachers have greater capacity to assess whether the questions generated by AIQUIZ are consistent with the subject or adequately cover the key points of course in the way they consider appropriate. Future work should include teacher interviews or analysis of

dashboard usage to measure the effectiveness of AIQUIZ in guiding instruction and supporting evidence-based teaching strategies.

For educational practice, the findings underscore the value of integrating LLM-powered tools such as AIQUIZ into teaching workflows. As an open-source platform, AIQUIZ is freely available for instructors to adopt and adapt within their own curricular contexts. Its core advantage lies in automating the labor-intensive task of MCQ creation, allowing educators to dedicate more time to direct student interaction, lesson planning, and other high-impact pedagogical activities.

In addition, the platform's teacher dashboard offers real-time, data-driven insights into student learning. By identifying persistent misconceptions or topic-specific struggles, such as difficulties with "CSS Selectors" or the "Unwind Operator" in MongoDB, instructors can tailor their instruction to address these gaps proactively. This promotes a shift toward evidence-informed teaching and more agile curriculum adaptation. However, the deployment of tools that log student interactions also raises important considerations around data privacy, institutional governance frameworks, and teacher training, all of which must be addressed for responsible adoption (Amo-Filvà et al., 2026).

Architecturally, AIQUIZ is scalable and adaptable to other educational domains through configuration files that directly impact the prompts that the app forwards to the LLM. However, its current operational success relies on using a prompt-based approach and the inherent knowledge of the LLM, making it immediately suitable for subjects like programming and databases where the model's training data is rich. Adapting AIQUIZ to domains less extensively covered by general LLMs would likely require more complex strategies, such as the future integration of Retrieval-Augmented Generation (RAG) using course-specific materials. It is therefore important to avoid overgeneralizing these findings beyond computing education. Domains with limited digital representation in LLM training corpora, highly localized knowledge, or greater conceptual ambiguity may yield substantially different results.

Looking ahead, several human–AI co-creation workflows could be explored to further enhance MCQ quality while reducing the validation burden for educators. The specific failure modes identified offer clear targets for future work. Given that 73.36% of confirmed errors were related to question formulation (e.g., ambiguous wording or poor distractors), future prompts should prioritize strict syntactic constraints. Furthermore, the finding that 35.37% of errors involved the LLM selecting an incorrect key suggests that model fine-tuning should focus specifically on logic verification rather than just content generation. One promising direction is to integrate an AI-driven critique or verification stage, in which a second LLM reviews each generated MCQ to detect issues such as multiple correct answers, incorrect keys, or poorly constructed distractors before questions reach students. Additionally, incorporating lightweight teacher review interfaces that highlight only AI-flagged "at-risk" items could substantially reduce manual effort while maintaining reliability. Therefore, the optimized co-creation workflow should shift the human validator's role from reviewing the entire question to a targeted verification of the answer key and the plausibility of the distractors, allowing educators to focus their effort where it has the highest pedagogical impact. Together, these hybrid workflows could strengthen quality assurance and move toward a more scalable, efficient model of human–AI co-creation in MCQ generation.

## 6 Conclusions

This study explored the use of AI-generated MCQs in higher computing education, using the cyclical model of self-regulated learning (SRL) as an analytical framework. Our investigation yields three primary findings. First, LLMs can generate topically relevant assessment content in real-world settings, as suggested by a student performance rate of 70.79% and confirmed by expert review. These data indicate utility and acceptability rather than demonstrating causal learning gains. Second, student perceptions of the tool are overwhelmingly positive; learners view the platform as useful and engaging, which is consistent with the motivational components of the SRL forethought phase. Third, while the system is reliable, our error analysis identified specific, recurring failure modes, such as incorrectly marked answers and plausible but inaccurate distractors, demonstrating that AI generation must currently be paired with human oversight.

**Implications for Research and Practice.** Theoretically, this work suggests that AI tools can support functionalities aligned with the phases of Zimmerman's SRL model: facilitating planning (forethought), enabling adaptive practice (performance control), and triggering error detection (self-reflection). Practically, the findings suggest that educators can leverage open-source platforms like AIQUIZ to massively scale the production of formative assessments, provided they implement lightweight verification processes for the generated answer keys. However, it is important to note that the empirical evidence collected measures student perception, usage, and acceptance rather than direct indicators of self-regulated learning strategy development. Future work incorporating validated SRL scales, longitudinal designs, or behavioral indicators of study strategy change would be needed to establish a causal link.

**Limitations.**

These conclusions should be interpreted in light of the following limitations. First, the measurement of question accuracy relied partially on student reporting, which may introduce false negatives: students may fail to identify or report flawed questions, potentially underestimating the actual error rate beyond the 0.85% calculated from reports. Compounding this, nearly 30% of student-reported errors were false positives, confirming that student feedback, while valuable, is not a consistently reliable measure of question quality. To estimate the magnitude of false negatives, future studies should incorporate random sampling of non-reported questions for expert review. Second, the study was conducted with undergraduate computing students at a single institution, limiting generalizability to other disciplines or educational levels. The high performance observed may be partially attributed to the abundance of computing-related content in the LLM's training corpus, an advantage that may not translate to less structured domains. Additionally, results are specific to GPT-4o Mini and the prompts used; different models or prompting strategies could yield different outcomes. Third, usage was not uniformly distributed across students. Future analyses should examine whether a small subset of highly active users disproportionately influenced aggregate metrics such as question generation counts and reported error rates.

**Future Work.** Future research should move beyond usage analysis to experimental designs, such as A/B testing, to isolate the specific learning gains attributable to AI-driven adaptive quizzing. Specifically, validated SRL scales (e.g., MSLQ) or longitudinal analysis of study behavior changes could provide direct evidence of self-regulation development, beyond the perception data gathered here. Furthermore, integrating learning

analytics with academic performance data could reveal whether these tools primarily benefit struggling students or accelerate those who are already high-performing.

Article submission date: April 18, 2026
Approval date: June 24, 2026
Publication date: July 1, 2026

Barra, E., Pilicita, A., Conde, J., Pozo, A., López-Pernas, S. & Reviriego, P. (2026). Using AI-powered multiple-choice question generation for self-regulated learning. *Revista de Educación a Distancia, 26* (84). http://dx.doi.org/10.6018/red.711091

**Declaración de las personas autoras sobre el uso de LLM**

Este artículo ha utilizado como referencia u objeto de investigación una herramienta LLM o el texto generado por ella, lo que se referencia en los términos que establece APA 7.0

**Declaración de las contribuciones de las personas autoras**

Conceptualización, E.B. y A.P.; curación de datos, A.P.; software, E.B.; análisis formal, J.C., S.LP. y A.P.; obtención de fondos, P.R.; investigación, E.B., A.P. y S.LP.; metodología, J.C. y E. B.; supervisión, P.R.; redacción – borrador original, E.B., A.P, J.C., A.P., y S.LP; redacción – revisión y edición, E.B., A.P, J.C., A.P., y S.LP.

**Financiación**

Este trabajo ha sido financiado por la Agencia Estatal de Investigación (AEI) 10.13039/501100011033 a través del proyecto FuN-4Date. También por el Grant PID2022-136684OB-C22, por la European Commission a través de Chips Act Joint Undertaking project SMARTY (Grant no. 101140087) y por TUCAN6-CM (TEC-2024/COM-460), financiado por CM (ORDEN 5696/2024).

**Ethical approval and informed consent statements**

This study was conducted in accordance with UPM ethical research guidelines. Prior to participation, students were informed that they were taking part in a research study evaluating the effectiveness of AI-generated multiple-choice questions in education. Participation was entirely voluntary, and students had the option to use the AIQUIZ platform without any obligation. No personally identifiable information was collected or shared with third parties. The study was designed to ensure minimal risk to participants, and all data were anonymized before analysis.

**Data availability statement**

All data generated or analysed during this study are included in this published article and its supplementary information files.

The datasets generated and analysed during the current study will be published in open access repositories such as Kaggle, HuggingFace or Github.

## Referencias

Amo-Filvà, D., Guàrdia Ortiz, L., Donate-Beby, B., Bautista Pérez, G., & Fanni, L. (2026). Integración de la Inteligencia Artificial y la Alfabetización de Datos en la ESO: Análisis de percepciones y condiciones de adopción. *Revista de Educación a Distancia (RED)*, *26*(83), 1–01. https://doi.org/10.6018/RED.690641

Badali, S., Rawson, K. A., & Dunlosky, J. (2023). How do Students Regulate Their Use of Multiple Choice Practice Tests? *Educational Psychology Review*, *35*(2), 1–26. https://doi.org/10.1007/S10648-023-09761-1/TABLES/4

Biancini, G., Ferrato, A., & Limongelli, C. (2024). Multiple-Choice Question Generation Using Large Language Models: Methodology and Educator Insights. *UMAP 2024 - Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*, 584–590. https://doi.org/10.1145/3631700.3665233

Cabellos, B., Rey, U., Carlos, J., Alcorcón, E., De Aldama, C., & Pozo, J. I. (2026). Creencias del alumnado de Formación Profesional sobre el uso de la inteligencia artificial generativa en la enseñanza y el aprendizaje. *Revista de Educación a Distancia (RED)*, *26*(83), 7–8. https://doi.org/10.6018/RED.671331

Ch, D. R., & Saha, S. K. (2020). Automatic Multiple Choice Question Generation from Text: A Survey. *IEEE Transactions on Learning Technologies*, *13*(1), 14–25. https://doi.org/10.1109/TLT.2018.2889100

Cheung, B. H. H., Lau, G. K. K., Wong, G. T. C., Lee, E. Y. P., Kulkarni, D., Seow, C. S., Wong, R., & Co, M. T. H. (2023). ChatGPT versus human in generating medical graduate exam multiple choice questions—A multinational prospective study (Hong Kong S.A.R., Singapore, Ireland, and the United Kingdom). *PLOS ONE*, *18*(8), e0290691. https://doi.org/10.1371/JOURNAL.PONE.0290691

Doughty, J., Wan, Z., Bompelli, A., Qayum, J., Wang, T., Zhang, J., Zheng, Y., Doyle, A., Sridhar, P., Agarwal, A., Bogart, C., Keylor, E., Kultur, C., Savelka, J., & Sakr, M. (2023). A Comparative Study of AI-Generated (GPT-4) and Human-crafted MCQs in Programming Education. *IFAC Symposium on Advances in Control Education*, 114–123. https://doi.org/10.1145/3636243.3636256

Gamage, S. H. P. W., Ayres, J. R., Behrend, M. B., & Smith, E. J. (2019). Optimising Moodle quizzes for online assessments. *International Journal of STEM Education*, *6*(1), 1–14. https://doi.org/10.1186/S40594-019-0181-4/FIGURES/11

Graham, M. J., Milanowski, A. T., & Miller, J. B. (2012). *Measuring and promoting inter-rater agreement of teacher and principal performance ratings.*

Grévisse, C., Pavlou, M. A. S., & Schneider, J. G. (2024). Docimological Quality Analysis of LLM-Generated Multiple Choice Questions in Computer Science and Medicine. *SN Computer Science*, *5*(5), 1–14. https://doi.org/ttps://doi.org/10.1007/s42979-024-02963-6

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied Measurement in Education*, *15*(3), 309–333. https://doi.org/10.1207/S15324818AME1503_5

Hang, C. N., Wei Tan, C., & Yu, P. D. (2024). MCQGen: A Large Language Model-Driven MCQ Generator for Personalized Learning. *IEEE Access*, *12*, 102261–102273. https://doi.org/10.1109/ACCESS.2024.3420709

Kıyak, Y. S., Coşkun, Ö., Budakoğlu, I. İ., & Uluoğlu, C. (2024). ChatGPT for generating multiple-choice questions: Evidence on the use of artificial intelligence in automatic item generation for a rational pharmacotherapy exam. *European Journal of Clinical Pharmacology*, *80*(5), 729–735. https://doi.org/10.1007/S00228-024-03649-X

Kumar, A. P., Nayak, A., K, M. S., Chaitanya, & Ghosh, K. (2024). A Novel Framework for the Generation of Multiple Choice Question Stems Using Semantic and Machine-Learning Techniques. *International Journal of Artificial Intelligence in Education*, *34*(2), 332–375. https://doi.org/10.1007/s40593-023-00333-6

Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2020). A Systematic Review of Automatic Question Generation for Educational Purposes. *International Journal of Artificial Intelligence in Education*, *30*(1), 121–204. https://doi.org/10.1007/S40593-019-00186-Y/TABLES/17

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159. https://doi.org/10.2307/2529310

Lehtinen, T., Haaranen, L., & Leinonen, J. (2023). Automated Questionnaires About Students' JavaScript Programs: Towards Gauging Novice Programming Processes. *ACM International Conference Proceeding Series*, 49–58. https://doi.org/https://doi.org/10.1145/3576123.3576129

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing System*, 9459–9474. https://doi.org/10.5555/3495724.3496517

Mistry, N. P., Saeed, H., Rafique, S., Le, T., Obaid, H., & Adams, S. J. (2024). Large Language Models as Tools to Generate Radiology Board-Style Multiple-Choice Questions. *Academic Radiology*, *31*(9), 3872–3878. https://doi.org/10.1016/J.ACRA.2024.06.046

Mulla, N., & Gharpure, P. (2023). Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications. *Progress in Artificial Intelligence*, *12*(1), 1–32. https://doi.org/10.1007/S13748-023-00295-9/TABLES/10

Purchase, H., Hamer, J., Denny, P., & Luxton-Reilly, A. (2010). The Quality of a PeerWise MCQ Repository. *Proceedings of the Twelfth Australasian Conference on Computing Education*. https://doi.org/10.5555/1862219.1862238

Qian, Y., & Lehman, J. (2017). Students' misconceptions and other difficulties in introductory programming: A literature review. *ACM Transactions on Computing Education*, *18*(1). https://doi.org/https://doi.org/10.1145/3077618

Shoaib, M., Husnain, G., Sayed, N., Yasin Ghadi, Y., Alajmi, M., & Qahmash, A. (2025). Automated Generation of Multiple-Choice Questions for Computer Science Education Using Conditional Generative Adversarial Networks. *IEEE Access*, *13*, 16697–16715. https://doi.org/10.1109/ACCESS.2025.3530474

Song, T., Tian, Q., Xiao, Y., & Liu, S. (2024). Automatic Generation of Multiple-Choice Questions for CS0 and CS1 Curricula Using Large Language Models. *Communications in Computer and Information Science*, *2023 CCIS*, 314–324. https://doi.org/10.1007/978-981-97-0730-0_28

Tek, F. B., Benli, K. S., & Deveci, E. (2018). Implicit Theories and Self-Efficacy in an Introductory Programming Course. *IEEE Transactions on Education*, *61*(3), 218–225. https://doi.org/10.1109/TE.2017.2789183

Tran, A., Angelikas, K., Rama, E., Okechukwu, C., Smith, D. H., & MacNeil, S. (2023). Generating Multiple Choice Questions for Computing Courses Using Large Language Models. *Proceedings - Frontiers in Education Conference, FIE*. https://doi.org/10.1109/FIE58773.2023.10342898

Vu, S. T., Truong, H. T., Do, O. T., Le, T. A., & Mai, T. T. (2024). A ChatGPT-based approach for questions generation in higher education. *AIQAM '*, *24*, 13–18. https://doi.org/10.1145/3643479.3662056

Wang, J., Xiao, R., & Tseng, Y.-J. (2025). Generating AI Literacy MCQs: A Multi-Agent LLM Approach. *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 2*, 1651–1652. https://doi.org/10.1145/3641555.3705189

Zimmerman, B. J. (2000). Attaining Self-Regulation: A Social Cognitive Perspective. *Handbook of Self-Regulation*, 13–39. https://doi.org/10.1016/B978-012109890-2/50031-7