UNIVERSITY OF PLOVDIV PAISII HILENDARSKI
FACULTY OF MATHEMATICS AND INFORMATICS
DEPARTMENT OF COMPUTER INFORMATICS

Ivan Marianov Ivanov

A MODEL AND SYSTEM FOR AUTOMATED GENERATION AND VERIFICATION OF TEST QUESTIONS USING ARTIFICIAL INTELLIGENCE

ABSTRACT

of a dissertation

for awarding the educational and scientific degree "Doctor"

Field of higher education: 4. Natural sciences, mathematics and informatics
Professional direction: 4.6. Informatics and Computer Science
Doctoral program: Informatics

Scientific Supervisors:
Prof. Stanka Hadzhikoleva, Ph.D.
Prof. Georgi Dimitrov, Ph.D.

Plovdiv, 2026

The dissertation was discussed and recommended for public defense by the Department of Computer Science at the Faculty of Mathematics and Informatics, University of Plovdiv Paisii Hilendarski, at a departmental meeting held on 9 July 2026.

The dissertation entitled “A Model and System for Automated Generation and Verification of Test Questions Using Artificial Intelligence” comprises 144 pages. The bibliography contains 89 references. The author's publications related to the dissertation topic include 6 papers.

The public defense of the dissertation will take place on ..... at ..... in the Conference Hall of the University of Plovdiv Paisii Hilendarski (New Building, 236 Bulgaria Blvd., Plovdiv).

The dissertation and the related defense materials are available for public inspection at the Dean's Office of the Faculty of Mathematics and Informatics, University of Plovdiv Paisii Hilendarski (New Building, 236 Bulgaria Blvd., Plovdiv), on working days from 8:30 a.m. to 5:00 p.m.

Author: Ivan Marianov Ivanov

Title: A Model and System for Automated Generation and Verification of Test Questions Using Artificial Intelligence

Table of contents

INTRODUCTION

The rapid development of artificial intelligence in recent years has had a significant impact on various areas of society, including education. Modern AI-based technologies provide new opportunities for the analysis, automation, personalization, and optimization of educational processes. Their potential is particularly significant in enhancing the development of educational content, supporting the organization and delivery of instruction, and improving the assessment of learners' knowledge and skills. Generative models and chatbots based on large language models (LLMs) facilitate the automatic generation of test questions, the analysis of assessment results, and the provision of personalized feedback to learners. These capabilities create favorable conditions for the development of a new generation of intelligent e-testing systems that improve the efficiency, adaptability, and overall quality of the assessment process.

The main objective of the dissertation research is to develop a model, architecture, and software implementation of an intelligent electronic testing platform that uses generative artificial intelligence for the automated creation, verification, management, and assessment of test content.

To achieve the main objective, the following ***tasks*** have been set:

Task 1: To study existing electronic testing systems and identify the main functional and non-functional requirements for an intelligent electronic testing system;

Task 2: To develop a conceptual model and architecture of an intelligent electronic testing platform, including mechanisms for the automated generation and verification of test questions through large language models;

Task 3: To develop a module for automated adaptive assessment of learners, based on modern psychometric approaches and artificial intelligence;

Task 4: To implement a software prototype based on the proposed model;

Task 5: To conduct experimental studies to evaluate the quality and applicability of the developed solution.

The main hypothesis of the present research is that the use of generative artificial intelligence within a specialized electronic testing platform will support teachers in creating and managing test content, improve the quality and reliability of test questions through automated verification, and provide more effective, objective, and personalized assessment of learners.

The problem addressed by the dissertation research is motivated by the need to develop intelligent solutions for electronic assessment that overcome the limitations of traditional test creation systems. Existing platforms often rely on predefined rules, templates, and a considerable amount of manual work in preparing test content. Despite the widespread adoption of generative language models, there is a lack of integrated solutions that simultaneously provide automated question generation, quality control through independent verification, adaptive testing, and results analysis. This necessitates the development of a new approach to e-assessment that combines the capabilities of generative artificial intelligence with established pedagogical principles of assessment validity, reliability, and adaptability.

The structure of the dissertation reflects the theoretical research conducted, the design of the conceptual model, the development of the software prototype, and the results of the experiments carried out. The dissertation consists of an introduction, four chapters, a conclusion, a list of references, a list of publications related to the topic, and one appendix illustrating the creation of questions by one generative model and their verification by a second one.

Chapter 1, Overview, presents a study of the possibilities for using artificial intelligence in education and the application of generative language models and chatbots in electronic testing. The evolution of test creation systems and classical approaches to the automated generation of test questions are examined. The main requirements for a modern intelligent electronic testing platform are formulated.

Chapter 2, Conceptual Model of an E-Testing Platform, proposes a conceptual model of an intelligent electronic testing system. The main user roles and functionalities, the platform architecture, the mechanisms for generating and verifying test questions through generative artificial intelligence, as well as some use cases, are presented. The interactions between the individual system components and the processes related to the creation, management, and assessment of test content in an intelligent educational environment are examined.

Chapter 3, Software Implementation, presents the implementation of the proposed platform, the technologies used, the system structure, and the main software modules. The mechanisms for integration with large language models, the management of test content, and the implemented tools for automated assessment are described. The architectural decisions, the development process, and the interaction between the individual system components are examined, as well as the implemented functionalities for generating, verifying, and administering tests in an electronic environment.

Chapter 4, Experiments, presents the results of the experimental studies conducted and the evaluation of the developed solution. The quality of the generated questions, the effectiveness of the verification mechanisms, and the applicability of the platform in a real educational environment are analyzed. The results of the experiments are also presented, confirming the effectiveness of the proposed approaches and the possibilities for their practical application in the process of electronic assessment.

The ***Conclusion*** provides an overview of the work carried out and the results achieved in relation to the defined objectives and tasks. The main scientific, scientific-applied, and applied contributions of the research are presented, confirming its relevance and practical significance. The prospects for the future development of the conducted research are also outlined.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my scientific supervisors, Prof. Dr. Stanka Hadzhikoleva and Prof. Dr. Georgi Dimitrov, for the opportunity to conduct the present dissertation research. I am grateful for the valuable advice, constructive recommendations, critical remarks, and guidance I received, which were of essential importance for achieving the set objectives and successfully completing the dissertation.

I would also like to express my deep appreciation to Dr. Eng. Todor Rachovski and Prof. Dr. Emil Hadzhikolev for their assistance, useful ideas, professional consultations, and support during the research. Their cooperation, shared experience, and constructive suggestions contributed significantly to the development of the present work and to overcoming a number of scientific and practical challenges.

CHAPTER 1. OVERVIEW

1.1. Artificial Intelligence in Education

In recent years, the rapid development of software technologies using artificial intelligence has created new opportunities for their integration into the educational process (Zhang & Aslan, 2021; Rojas & Chiappe, 2024). This includes the use of natural language processing technologies through large language models (Kasneci, Sessler et al., 2023; Jeon & Lee, 2023), image recognition technologies through computer vision (Chen, Samuel et al., 2021; Agbo et al., 2020), data analysis and forecasting through neural networks (Kucak, Juricic et al., 2018; Hadzhikolev, Hadzhikoleva et al., 2021), and others. The results of numerous studies show that AI has the potential to improve the quality of education, make it more accessible, and better align it with the individual needs of learners (Sofianos, Stefanidakis et al., 2024; Harry & Sayudin, 2023). Artificial intelligence is applied in various fields of education, including the social sciences (Nurhayati & Halimah, 2024), engineering education (Nunez & Lantada, 2020), natural sciences (Darayseh, 2023; Hadzhikoleva, Borisova et al., 2025), life sciences (Kandlhofer, Steinbauer et al., 2016), medical education (Briganti & Le Moine, 2020), foreign language learning (Edmett, Ichaporia et al., 2023), and many other educational domains (Hajkowicz, Sanderson et al., 2023). This indicates that AI is gradually becoming a universal tool for supporting learning and assessment, regardless of the specific subject area.

AI is particularly widely applied in STEM education. It is used to develop intelligent tutoring systems, profiling and prediction systems, as well as adaptive learning and personalization of educational content (Xu & Ouyang, 2022). Previous studies have demonstrated the positive role of AI in improving personalized learning experiences, identifying learners at risk, and automating administrative activities (Rahiman & Kodikal, 2024). In addition, various artificial intelligence algorithms are used for data analysis and quality assurance in higher education through the early detection of problems and deviations in key performance indicators (Mishra, 2019).

The development of large language models (Large Language Models – LLMs) marks the beginning of a new stage in the development of artificial intelligence and its application in education. In recent years, particular attention has been drawn to chatbots based on large language models. Among the most popular representatives of this class of systems are ChatGPT, Gemini, Claude, Copilot, and other similar solutions. They enable interactive communication with users through natural language and can perform a wide range of tasks related to learning, teaching, and research activities (Dempere, Modugu et al., 2023). Various studies confirm that generative chatbots, when used appropriately, can successfully support learners in acquiring new knowledge, solving tasks, preparing projects, and conducting independent research (Chaudhry, Sarwary et al., 2023; Pradana, Elisa et al., 2023). They can

also be used to create gamified resources (Borisova, Hadzhikoleva et al., 2024; Hadzhikoleva, Gorgorova et al., 2024).

1.2. Automated Assessment and Electronic Testing

The use of artificial intelligence in the assessment of learning enables the automation of a wide range of tasks that have traditionally been performed by instructors. These include the generation and evaluation of multiple-choice test questions, the analysis of free-text responses, the assessment of student assignments, and the provision of personalized feedback to learners. Research has shown that AI technologies can effectively support the assessment of both lower-order cognitive knowledge and higher-order skills, including analysis, interpretation, and problem-solving (Hadzhikolev et al., 2021).

Despite the significant advantages of automated assessment, its implementation is associated with a number of challenges. These include ensuring the reliability, objectivity, and transparency of assessment processes, as well as maintaining effective quality control over AI-generated content. The use of generative models also introduces additional challenges, including the risk of generating factually inaccurate content and the need for expert validation of automatically generated test questions.

1.3. Evolution of Test Creation Systems

The creation of tests usually requires significant time and human resources. In addition, tests must be evaluated for reliability and validity, and pilot testing is essential in order to identify weaknesses and omissions and to eliminate them (Osterlind, 1998). This motivates teachers to actively seek ways to automate this process, including the generation of test questions, the construction of tests, and their verification (Bugbee, 1996). Initially, applications using traditional methods and technologies were developed.

Classical approaches to test creation have several significant ***limitations***:

- Traditional systems are ***highly dependent on predefined rules and templates***, which limits their ability to generate diverse questions, especially when more complex or adaptive test formats are required.
- These systems require ***considerable manual configuration*** both in creating rules and in selecting appropriate keys and distractors, which makes them time-consuming and labor-intensive.
- ***Adaptation is limited***; it is usually based on previous responses and cannot easily adjust to dynamically changing conditions.
- ***The applications use limited semantic information processing***, which reduces their ability to understand and interpret deeper contexts and relationships in the text.

Generative artificial intelligence models provide an opportunity to rapidly create a large number of diverse questions with different levels of complexity and to reduce the time required for test preparation. Despite the considerable capabilities of modern generative models, the quality of the generated questions depends to a significant extent on the quality of the instructions provided. Unlike traditional software systems, in which the user works through a predefined interface, interaction with large language models is carried out through text queries, or prompts. In order to obtain correct, pedagogically sound test questions aligned with the learning objectives, the teacher must formulate a sufficiently detailed and structured prompt.

The basic information that should be included in such a prompt comprises:

- ***Topic or learning content*** – defines the subject area and the knowledge to be assessed. The more specifically it is defined, the more relevant the generated questions are.
- ***Number of questions*** – determines the volume of the generated content.
- ***Type of questions*** – the type of questions must be specified, such as single-answer, multiple-choice, true/false, open-ended questions, etc., since different formats assess different aspects of knowledge.
- ***Difficulty level*** – specifying the level allows the questions to be aligned with the learners' preparation and prevents the generation of tasks that are either too easy or too complex.
- ***Educational level or target audience*** – the content and terminology of the questions must be aligned with the audience, such as school pupils, university students, specialists, and others.
- ***Structure of the answers and correct answer***. In closed-ended questions, the model must generate not only the question stem but also a set of possible answers, among which the correct one must be clearly identified. Well-constructed answers are a key factor for the reliability and validity of test-based assessment.

From the perspective of pedagogical design, the most critical elements are the topic, the question type, the difficulty level, and the educational level, because they determine whether the generated questions will be appropriate for the specific learning objective at all (Fig. 1).

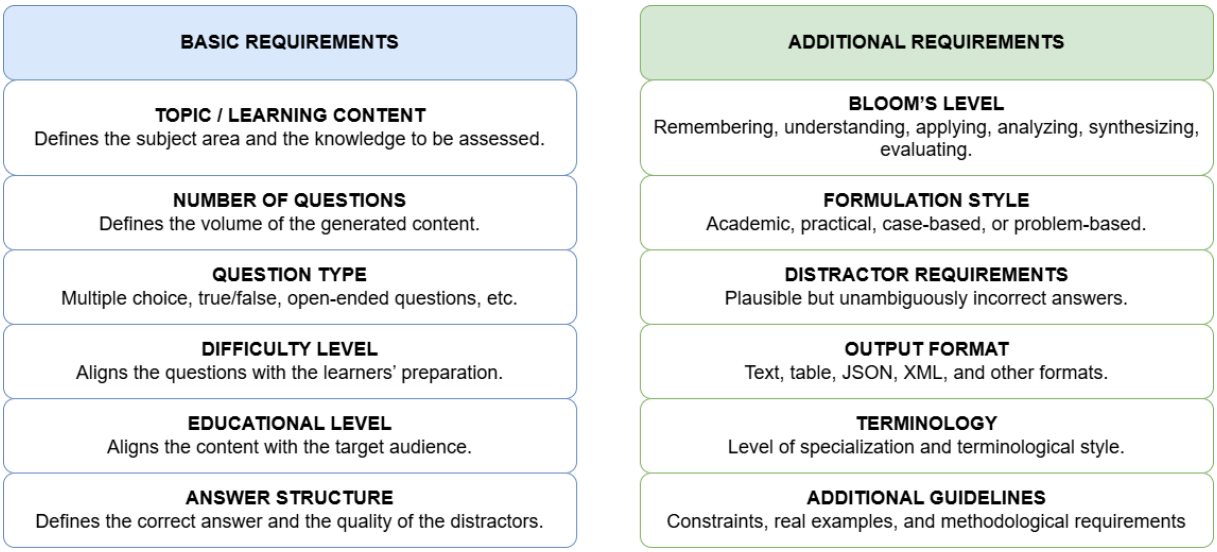

Figure 1. Requirements for a Good Prompt for Generating Test Questions

Additional requirements may also be included in the prompt in order to improve the quality and pedagogical value of the test questions. These may include, for example:

- ***Cognitive level according to Bloom's taxonomy***. Specifying the cognitive level allows the questions to be directed toward a particular type of cognitive activity. Depending on the learning objectives, the questions may assess the memorization of facts, understanding of concepts, application of knowledge in

new situations, analysis of relationships, evaluation of decisions, or creation of new ideas and approaches.

- ***Formulation style***. Questions can be generated in different styles according to the specifics of the academic discipline. Possible styles include an academic style, practically oriented questions, case-study-based questions, problem-based tasks, or questions related to real situations from professional practice.
- ***Requirements for distractors***. The quality of incorrect answers is essential for the reliability of the test. Well-formulated distractors should be plausible and thematically related to the question, without being misleadingly ambiguous or obviously incorrect.
- ***Output format***. The generated content may be presented in different formats depending on the method of subsequent processing, such as plain text, a table, JSON, XML, or another structured format suitable for use.
- ***Terminology used***. A preferred level of specialization and terminological style may be specified. For example, the questions may use generally accessible terminology for beginner learners or specialized professional terminology for advanced students and experts.
- ***Additional constraints and guidelines***. The prompt may include specific requirements regarding the formulation of the questions, such as avoiding negative constructions, preventing misleading or ambiguous statements, using real examples, including practical situations, or complying with specific educational standards and methodological recommendations.

1.4. Basic Requirements for an Intelligent Electronic Testing Platform

A modern electronic testing platform should meet a set of functional and non-functional requirements that ensure efficiency, reliability, and pedagogical value.

First, the system must provide ***capabilities for creating and managing test content***. This includes support for various types of test questions, such as multiple-choice questions, open-ended questions, matching questions, and others. Modern solutions should support the automated generation of questions through the use of generative models, with criteria such as topic, difficulty level, cognitive level, and others being specified.

Another important requirement is the ***availability of mechanisms for validating and editing the generated content***. The teacher must be able to review, correct, and approve test questions before they are used in order to exercise control over their quality. This is particularly important when AI-based generative systems are used, where there is a risk of incorrect or inappropriate content.

Another key functionality is the ***management of tests and test sessions***. The system must allow the configuration of parameters such as the number of questions, the order in which questions are presented, time limits, access conditions, and others. In this context, it is important to support both static tests and adaptive tests, in which the sequence of questions is determined dynamically depending on the learner's performance.

Modern systems should provide ***automatic assessment of answers***, as well as the possibility of manual review when necessary. For closed-ended questions, assessment can be

fully automated, while for open-ended questions, mechanisms should be provided to support assessment, including through the use of artificial intelligence.

A particularly important requirement is ***support for adaptive testing***. This involves the use of psychometric models, such as Item Response Theory, to dynamically determine the next question based on the estimated ability of the learner. In this way, more precise and personalized assessment can be achieved with a smaller number of questions.

Another important aspect is the reliability of the AI models used. Different models may interpret the same task in different ways and generate content of varying quality. For this reason, it is advisable to implement a mechanism for ***multi-model verification*** in the system, in which questions generated by one language model are evaluated by another independent model.

1.5. Conclusions

The studies conducted show that generative models can be successfully used for creating test questions, generating learning materials, and providing feedback to learners. However, there are a number of challenges related to the quality, reliability, and pedagogical suitability of the generated content.

Based on the analysis carried out, the aim and tasks of the present dissertation have been formulated and presented in the Introduction.

The implementation of the defined tasks will create the prerequisites for building a reliable and intelligent electronic testing environment that supports teachers in creating test content and ensures more effective, objective, and personalized assessment of learners.

CHAPTER 2. CONCEPTUAL MODEL OF AN E-TESTING PLATFORM

2.1. Main Roles and Functionalities

The conceptual model of the electronic testing platform defines ***three main user roles***: administrator, teacher, and learner. Each role has access to specific functionalities.

The ***administrator*** is responsible for the overall management and maintenance of the system. The administrator's activities include user management, access control, configuration of basic system parameters, activity monitoring, and log analysis. The administrator also plays a key role in managing the AI functionalities by configuring the models used, defining policies and restrictions for their use, and monitoring the workload, efficiency, and costs associated with AI services. In addition, the administrator is responsible for security, report generation, as well as data backup and recovery.

The ***teacher*** manages the main processes related to the creation, administration, and analysis of tests. The teacher creates test questions manually or through AI-based generation, after which they review, correct, validate, and approve them. Within the testing process, the teacher configures tests, defines their structure and content, assigns them to learners, and monitors their completion. After the tests are completed, the teacher analyzes the results using assessment and statistical tools and provides feedback. Artificial intelligence supports the generation and verification of questions, the analysis of results, and the formulation of personalized feedback.

The ***learner*** is the main user of the electronic testing system. The learner participates in test sessions assigned by the teacher, as well as in adaptive tests for self-study. In adaptive testing, the system generates a sequence of questions aligned with the learner's current level

of knowledge, which supports more precise assessment. After completing the tests, the learner has access to their results, test history, and performance statistics. The system can provide AI-based feedback, including explanations of correct answers, analysis of mistakes, and recommendations for improvement.

2.2. General Architecture

The electronic testing platform is based on a multi-layer architecture. It ensures a clear separation of responsibilities, flexibility, and scalability. In general, the architecture can be considered as consisting of four main layers (Fig. 2). These include:

- User Layer;
- Application Layer;
- AI Layer;
- Data Layer.

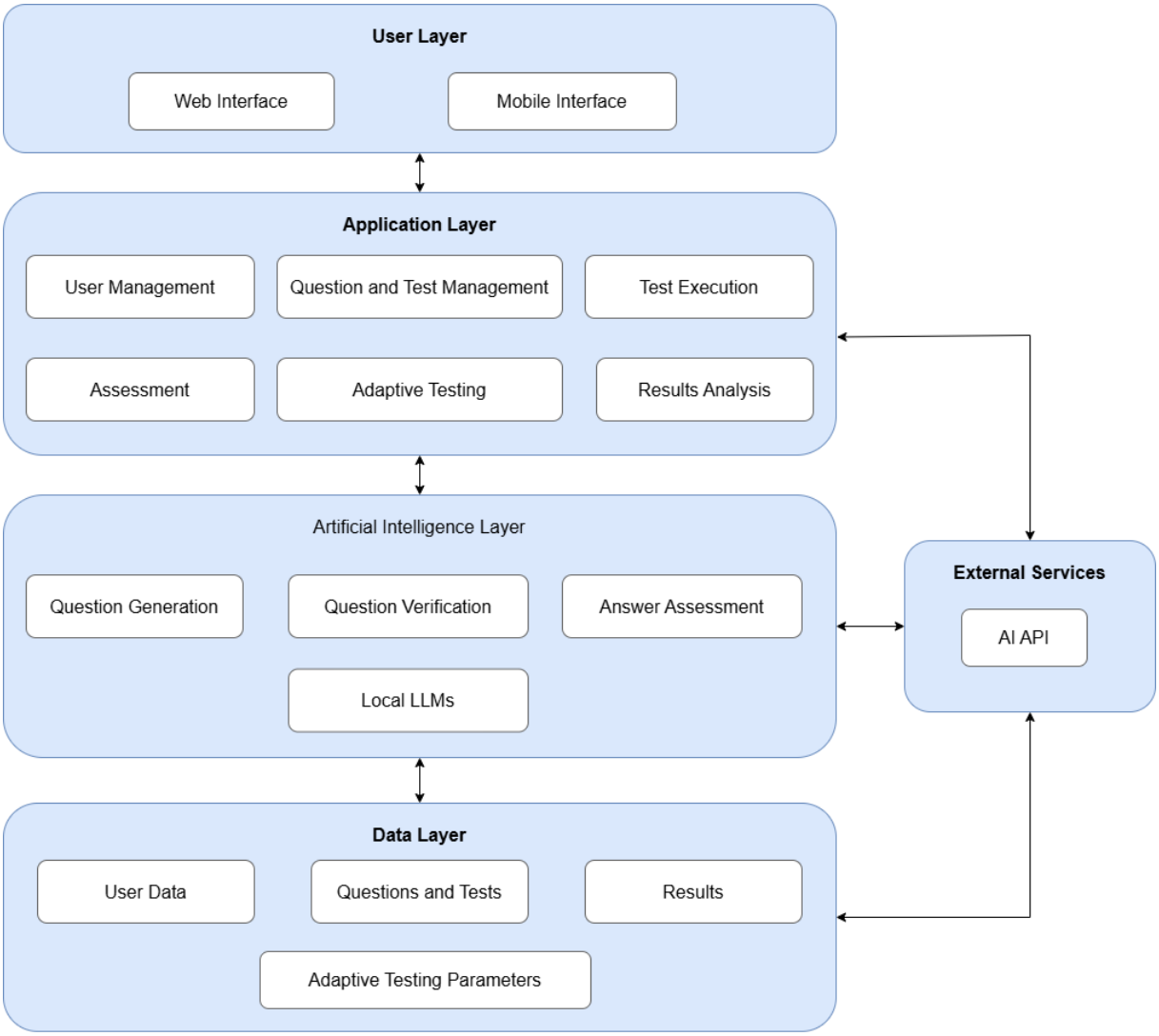

Figure 2. System Architecture

The ***User Layer*** provides an interface for interacting with the system through web or mobile applications. It provides access to the platform functionalities depending on the user's role, including test visualization, tools for creating and configuring questions, presentation of

results and analyses, and others. The main purpose of this layer is to ensure intuitive and efficient access to the system, regardless of the users' technical background.

The ***Application Layer*** implements the main business logic of the system and coordinates the interaction between the other layers. It includes modules for user management, question and test management, execution of test sessions, assessment, and analysis of results. This layer also implements the mechanisms for adaptive testing, which use the learner's current results to dynamically determine the next questions. The Application Layer acts as an intermediary between the user interface and the intelligent services, ensuring consistency and control over the processes.

The ***AI Layer*** is a key component of the architecture that provides the intelligence of the system. It includes integration with large language models, which are used for the automated generation of test questions, as well as for their verification and evaluation. This layer implements mechanisms for request management, or prompt engineering, processing of results in a structured format, such as JSON, and combining the results from different models in order to increase reliability. Through abstraction of the AI functionalities, it becomes possible to easily replace or add new models without changing the other parts of the system.

The ***Data Layer*** ensures the storage and management of all data in the system. It includes a database for users, questions, tests, test results, and parameters related to adaptive assessment. The data are used both for the current operation of the system and for subsequent analysis and improvement of test quality. In addition, this layer can support the storage of historical data required for calibrating the parameters used in adaptive testing.

2.3. Main Modules

The ***Question Generation Module*** uses large language models for the automated creation of test content. The input data to the module include criteria specified by the teacher, such as topic, difficulty level, cognitive level according to Bloom's taxonomy, question type, and additional instructions. Based on these parameters, a request is constructed and sent to the language model, which generates structured questions and answers. The result is presented in a standardized format, allowing subsequent processing and integration with the system.

In order to increase the reliability of the generated content, the system includes a ***Verification Module***, which uses an additional AI model to assess the quality of the questions. This module analyzes correctness, clarity, compliance with the specified criteria, and potential problems such as ambiguity or errors. More than one model may be used, which enables multi-model verification, with the results being combined to achieve higher accuracy. The module may also provide a confidence score that supports the teacher in decision-making (Fig. 3).

The ***Question and Test Management Module*** provides functionalities for creating, editing, storing, and organizing questions and tests. The teacher can select questions from a database or use AI-generated questions, combining them into tests in different ways. The module supports the configuration of parameters such as time limits, difficulty level, number of questions, and assessment criteria. It also provides the possibility of reusing questions, which increases the efficiency of the process.

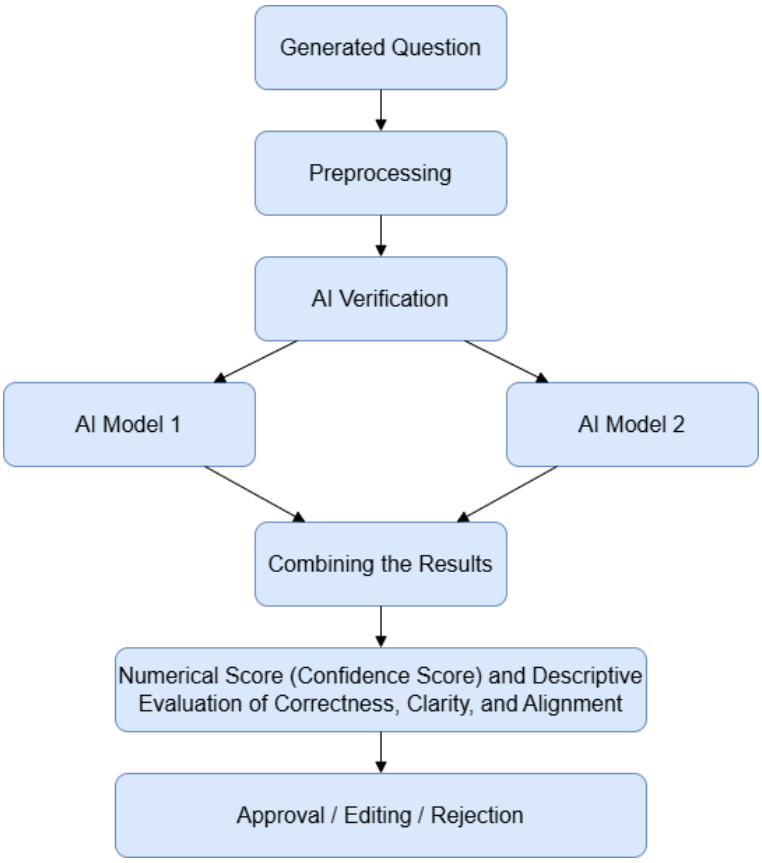

Figure 3. Process of Verification and Evaluation of Test Questions

The ***Adaptive Testing Module*** implements mechanisms for the dynamic selection of questions aligned with the learner's current level of knowledge. To implement adaptivity, Item Response Theory (IRT) is used as a psychometric model describing the relationship between the learner's latent ability and the probability of correctly answering a particular test question.

Each question is characterized by a set of psychometric parameters, such as discrimination, difficulty, and probability of guessing. Based on these parameters and the current estimate of the learner's ability (θ), the probability of a correct answer and the information value of each question from the test bank are calculated. The next question is selected so as to provide maximum information about the current ability estimate, which allows the standard error of measurement (SE) to be gradually reduced and the accuracy of the final assessment to be improved (Embretson & Reise, 2025; Baker & Kim, 2004; Wyse, 2010; Van der Linden & Glas, 2010; Wainer et al., 2000).

There are different IRT models, including 1PL, 2PL, and 3PL. The main difference between them lies in the number of parameters, that is, the characteristics of the question, taken into account by the model.

The probability $P(X = 1 \mid \theta)$that a learner with latent ability θ will answer a question correctly is calculated in the different IRT models using the following functions:

1PL Model: $P(X = 1 \mid \theta) = \frac{1}{1+e^{-(\theta-b)}}$

2PL Model: $P(X = 1 \mid \theta) = \frac{1}{1+e^{-a(\theta-b)}}$

3PL Model: $P(X = 1 \mid \theta) = c + \frac{1-c}{1+e^{-a(\theta-b)}}$,

where ***a*** is ***discrimination***, ***b*** is ***difficulty***, and ***c*** is ***the probability of guessing the correct answer*** to the question.

Adaptive tests are suitable for self-assessment, personalized learning, and experimental studies.

The ***Assessment Module*** processes learners' answers and calculates the results. In closed-ended questions, assessment is performed automatically, whereas in open-ended questions, AI-based methods may be used to support assessment, or manual review by the teacher may be applied. In the context of adaptive testing, the module also calculates and updates the value of the learner's latent ability.

The ***Analysis Module*** provides tools for processing and visualizing test results. It generates statistics on learners' performance, analytical parameters of the questions, such as difficulty and discrimination, as well as aggregated reports for teachers and administrators. The data can be used to improve the quality of tests and optimize the learning process.

The ***Artificial Intelligence Management Module*** acts as an intermediary between the application layer and external AI services. It manages communication with the language models, processes input and output data, and ensures the standardization of requests by constructing a unified prompt using the parameters specified by the teacher. In addition, the module may implement mechanisms for model selection and tracking the quality of the generated results.

2.4. Use Cases

The main use cases illustrating the interaction between the different roles and modules in the system are:

- Creation, including generation and verification, of test questions;
- Creation and configuration of a test;
- Completion of a test by a learner;
- Assessment and provision of feedback;
- Analysis of results and optimization.

2.5. Artificial Intelligence Integration

In the proposed model (Fig. 4), interaction with AI services is implemented through a specialized intermediate layer, the AI Integration and Management Layer, which provides a unified interface between the application logic of the system and the language models used. This layer is responsible for constructing requests, or prompts, based on the parameters specified by the user, selecting an appropriate model, processing and standardizing the obtained results, and managing communication with external AI services.

Through this architecture, the various functionalities related to artificial intelligence, including the generation of test questions, verification through multiple models, and support for assessment, use a common mechanism for interaction with LLMs, without requiring the application modules to be tied to a specific provider or technology.

For the generation of test questions, LLMs create content based on predefined parameters specified by the teacher. These may include the topic, difficulty level, cognitive level according to Bloom's taxonomy, question type, and additional instructions regarding the formulation style.

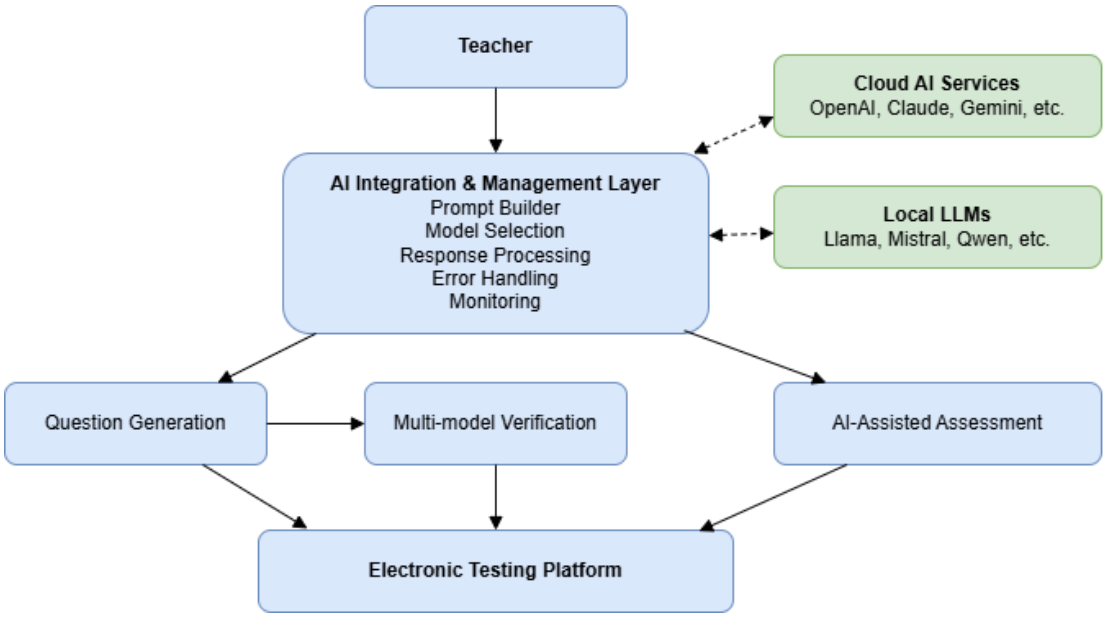

Figure 4. Architecture of Artificial Intelligence Integration

Interaction with the LLMs is carried out through specialized requests, or prompts, which are dynamically constructed by the AI Integration and Management Layer. The obtained results are processed, verified, and transformed into a unified structured format, such as JSON, which enables their automatic integration into the other components of the platform.

The obtained results are presented to the teacher in the form of recommendations and reliability indicators, while the final decision to accept, edit, or reject a given question remains entirely the responsibility of the teacher.

Artificial intelligence can also be used to support the assessment process, especially for open-ended questions. LLMs can analyze textual answers, compare them with expected solutions, and provide a preliminary assessment or recommendations. This enables partial automation of assessment and reduces the time required for checking tests.

However, the application of AI in assessment requires careful control, with the possibility of manual intervention and final evaluation by the teacher. In this way, automation is combined with human expertise, ensuring higher reliability.

2.6. Conclusions

The proposed conceptual model of a next-generation electronic testing platform aims to integrate the capabilities of generative artificial intelligence and adaptive assessment methods into a unified, modular, and extensible architecture. The model builds upon traditional electronic testing systems by providing a higher degree of automation, flexibility, and personalization of the assessment process.

CHAPTER 3. SOFTWARE IMPLEMENTATION

3.1. Technology Stack

Multiple technologies were used for the development of the platform, which, according to their purpose, are grouped into five categories: backend, databases, frontend, artificial intelligence providers, and infrastructure. The framework used is .NET 10.0. The technology stack is illustrated in Fig. 5.

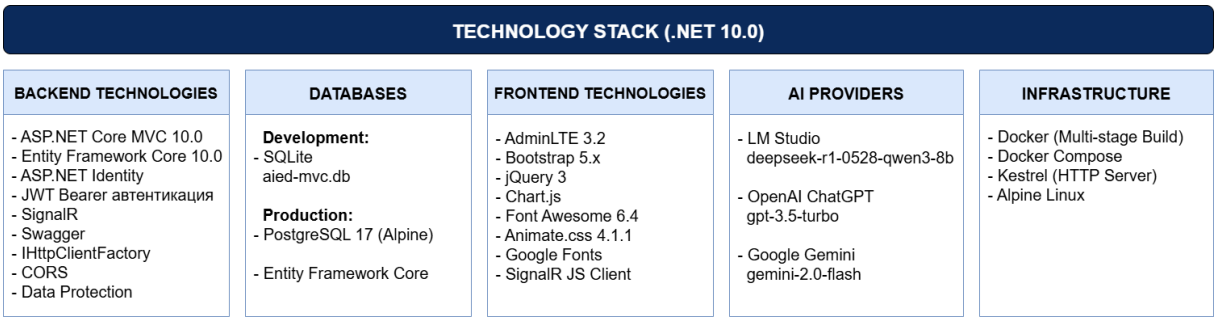

Figure 5. Technology Stack of the Application

3.2. Data Model

The database contains a set of main and auxiliary tables that support user management, questions, tests, adaptive assessment, AI generation, verification, and system configuration. MySQL Workbench was used to visualize the database structure and the relationships between the individual tables. Due to the large number of tables and references, the model is divided into several logical groups. For better readability and clarity of presentation, the relationships between the individual groups are not shown in the following diagrams.

User (Fig. 6) is a central model for the users of the system, including administrators, teachers, and learners. It extends the standard IdentityUser by adding fields for first name, last name, faculty, faculty number, which is unique, year of study, group, specialty, role, active status, and an additional permission to view all tests.

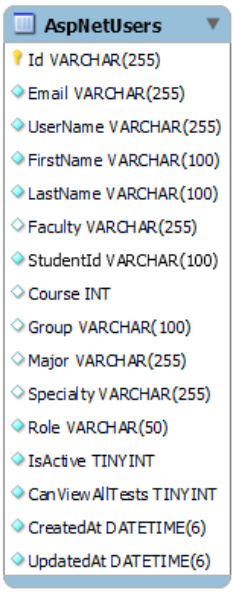

Figure 6. Model User

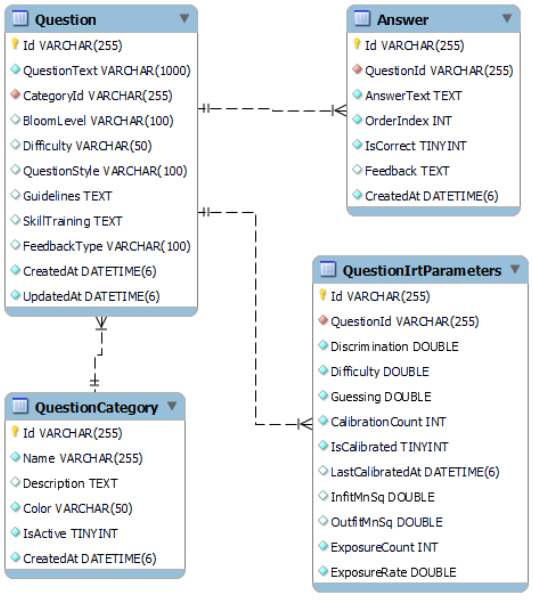

Figure 7. Model Question

The *Question* model (Fig. 7) represents a test question. It has the following characteristics: text of up to 1000 characters, category, Bloom's taxonomy level, difficulty, style, and additional guidelines. Each question has multiple Answer records with text, order, and a marker for the correct answer, as well as one *QuestionIrtParameters* record containing IRT parameters: discrimination (a), difficulty (b), guessing (c), *Infit* and *Outfit MnSq* statistics, and a calibration flag.

The *Test* model (Fig. 8) describes a test. Classical and adaptive tests have the following characteristics: title, description, category, creator, time limit, and flags indicating activity and public visibility. For adaptive tests, additional fields specify the maximum number of questions, the standard error threshold for terminating the test, and the category of the question pool. The relationship between a test and its questions is implemented through TestQuestion, which also defines their order.

TestAssignment (Fig. 8) models the assignment of a test. Its characteristics define the teacher assigning the test, the learners to whom the test is assigned, an optional deadline for completing the test, and notes. *TestSubmission* records the student's completed attempt, including the following information: score as a percentage, total number of answers and number of correct answers, status, time taken to complete the test, grade manually entered by the teacher on the six-point grading scale, comment, θ estimate from IRT, and standard error. StudentAnswer stores the answers given by the learner to each individual question.

The *AdaptiveTestSession* model tracks the state of an adaptive test while it is being completed by the learner. The *SystemSetting* model implements a mechanism for storing configuration parameters using a key-value scheme. It is used to manage global platform settings, such as the configuration of AI providers, adaptive testing parameters, validation thresholds, and others. The *GenerationHistory*, *QuestionValidationResult*, and *ErrorLog* models are used to store information about AI-based question generation and verification, as well as to record system errors and diagnostic events that support platform monitoring and maintenance.

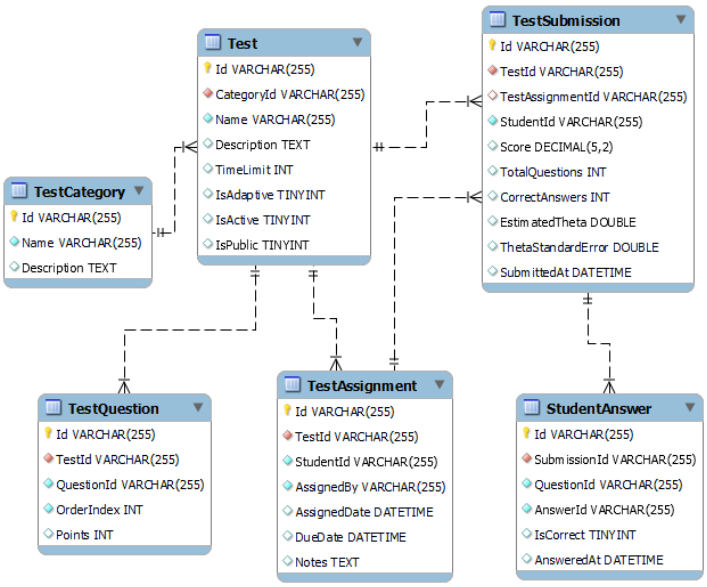

Figure 8. Test and TestAssignment Models

3.3. Main Functionalities

Teachers and administrators can create test questions in two ways: through manual entry or by generating them with the help of AI.

Manual creation is performed in two steps. In the first step, the teacher must enter the text of the question and metadata, including question category, difficulty level (Easy, Medium, Difficult, Very Difficult), question type ("multiple choice", "true/false", "essay", "short

answer"), and points. In the second step, additional information about the question is entered, depending on its type and Bloom's taxonomy level (Knowledge, Understanding, Application, Analysis, Synthesis, Evaluation). For example, if the "multiple-choice" type is selected, the possible answers must be added in the second step, with exactly one answer marked as correct.

When ***generating a question with AI***, in addition to category, difficulty level, question type, and Bloom's level, the teacher must specify:

- learning content, such as a file, text, or topic title, on the basis of which the questions are to be generated;
- the AI model to be used. Currently, the supported options are the local model deepseek-r1:8b and the cloud models GPT-5.4 mini by OpenAI, Gemini 2.5 Pro by Google, and Claude Sonnet 4.6 by Anthropic.

Optionally, additional parameters may be specified, including:

- additional instructions for the questions, such as a focus on practical examples, numerical problems, etc.;
- a second generative model to be used for evaluating the generated test questions.

These options are illustrated in Fig. 9.

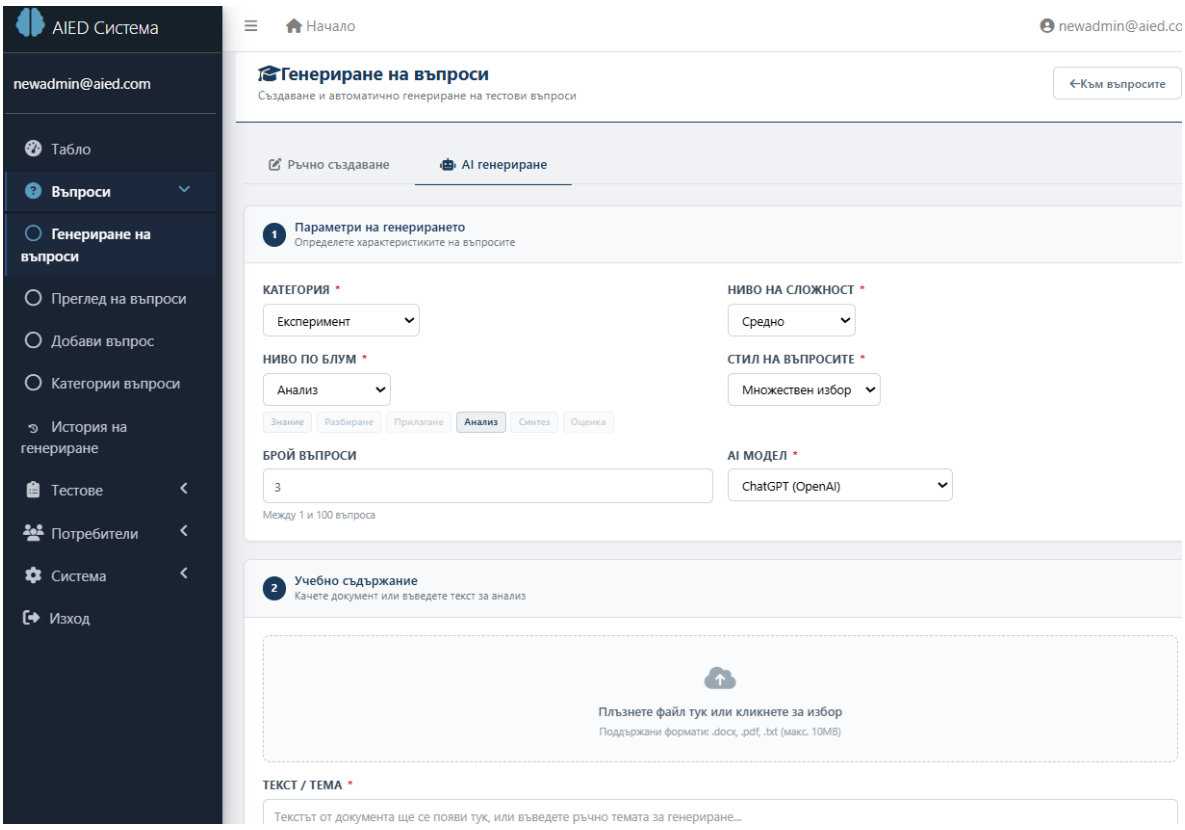

Figure 9. Screen for Automatic Generation of Test Questions by an LLM

After pressing the "Generate Questions" button, the system sends a prompt in Bulgarian to the selected AI provider and processes the JSON response.

If the teacher has requested evaluation of the questions by a generative model, a short evaluation and a link to an "Evaluation Questionnaire" containing the corresponding scores are displayed under each question.

Each of the generated questions can optionally be edited by the teacher and saved in the database under a specific question category.

The platform supports ***two types of tests***: classical and adaptive.

Classical tests are created by selecting questions from those previously created. During their configuration, the test name, description, category, difficulty level, and duration are specified. The test may have an active status, for assessment purposes, or public access, for self-study.

Adaptive tests are based on IRT and operate through dynamic selection of questions from a pool. When configuring such a test, the following parameters are specified: name, description, test category, question pool, defined by category, maximum number of questions, from 5 to 100, test duration, and stopping threshold for the standard error, from 0.1 to 1.0.

The teacher can assign a test to students selected from the list of students registered in the system. Optionally, a deadline for completing the test and explanatory notes may be specified (Fig. 10).

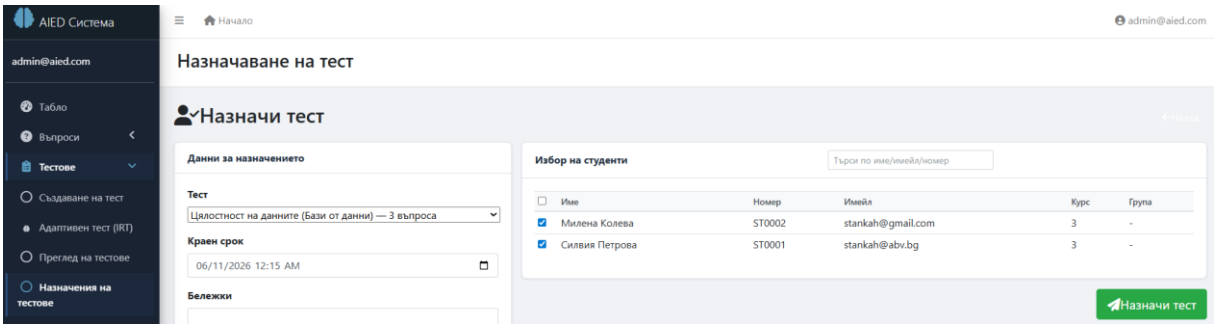

Figure 10. Test Assignment

In a ***classical test***, the student sees all questions at once, selects answers, and submits the test at the end. The score is calculated automatically based on the percentage of correct answers. After the test is submitted, a notification is sent via SignalR, which updates the teacher's monitoring screen.

When an ***adaptive test*** is started, the system creates a new test session and initializes the initial estimate of the learner's ability (θ). If a previous estimate exists for the learner, it is used as the initial value; otherwise, θ is initialized with a value of 0.0, corresponding to the average ability level. After each answer is given, the parameters used are updated. The test ends when:

- the target standard error is reached, namely the stopping SE threshold;
- the maximum number of questions is reached;
- the questions from the pool are exhausted;
- the test time expires.

When the test is finalized, a submission record is created in the database, including the values of θ, standard error, and percentage score of correct answers.

Automatic assessment calculates the score as the percentage of correct answers. Assessment is performed according to the six-point grading scale used in Bulgaria, with the following grading criteria adopted: Excellent 6 (90–100%), Very Good 5 (75–89%), Good 4 (60–74%), Satisfactory 3 (45–59%), and Poor 2 (0–44%).

The teacher can manually assign grades, which take priority over the automatically calculated ones, and add individual feedback comments.

The platform provides ***several types of reports***: a general report with the number of tests, questions, students, and average grade; test-level analysis with grade distribution; individual student performance with completed tests and trends; IRT analysis with calibrated parameters and average θ; and a detailed IRT report by question, including discrimination, difficulty, guessing, Infit/Outfit MnSq, and the ICC curve. General reports are available to administrators and teachers, while IRT analysis is available only to administrators.

3.4. Real-Time Monitoring with SignalR

The platform supports real-time monitoring, through which the teacher can track test completion and students' progress. SignalR is used for bidirectional communication between the server and the client applications. Whenever possible, it operates through WebSocket, and, if necessary, automatically switches to alternative mechanisms such as Long Polling or Server-Sent Events.

When the monitoring page is opened, the client application connects to the SignalR Hub and joins a group associated with the specific test. In this way, events related to a given test are sent only to the teachers who are monitoring it. When a test is submitted or an answer is provided in an adaptive test, the server sends events containing data about the student, the result, the submission time, the current ability estimate θ, the standard error, and the number of answered questions.

The interface visualizes both summary and detailed information, including the number of submitted tests, the average result, the number of active students, and data for each participant. The information is updated automatically without reloading the page. To ensure higher reliability, automatic reconnection is provided in the event of temporary interruptions, and the current connection status is displayed by means of a color indicator.

Figure 11 shows a screen for real-time test monitoring through SignalR. The upper part displays the main indicators of the current state of the test, while the central part presents a list of learners, their status, achieved points, and last activity.

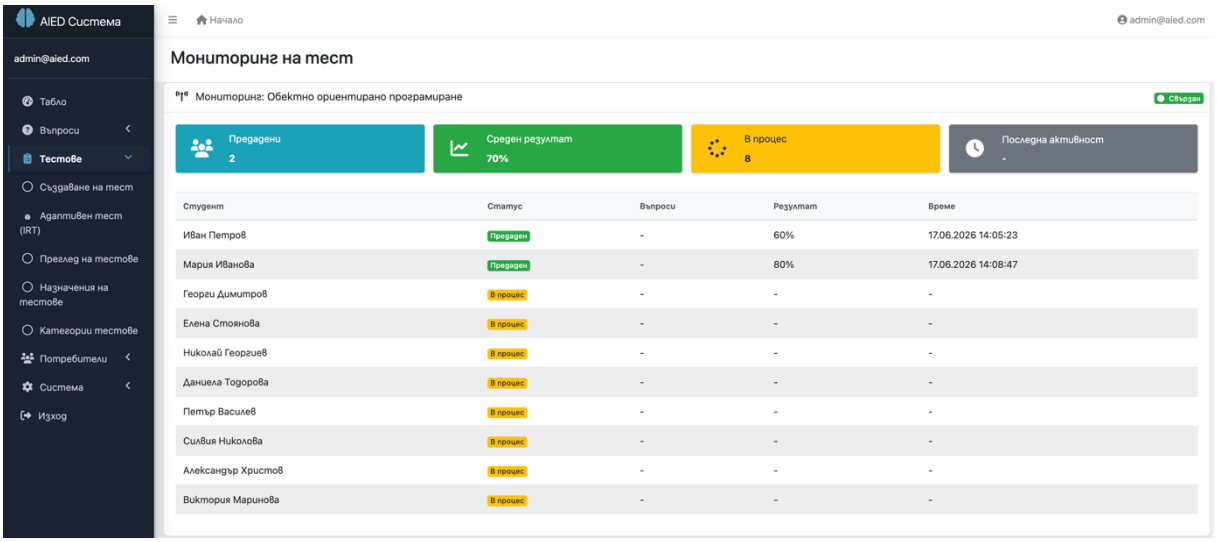

Figure 11. Test Monitoring During Execution Through SignalR

The main advantage of this approach is the automatic updating of data when server-side events occur, which allows the teacher to monitor learners' progress and promptly identify technical problems or unusual delays.

3.5. Conclusions

The intelligent electronic testing platform presented in this chapter implements the main functionalities of the conceptual model proposed in Chapter 2 and confirms its practical applicability.

CHAPTER 4. EXPERIMENTS

A pilot experimental study was conducted to evaluate the capabilities of generative artificial intelligence for the automated generation and evaluation of test questions. Within the study, the behavior of two large language models was analyzed, used in two different roles: as ***generators of test questions (GenModel)*** and as ***evaluators of their quality (EvalModel)***.

4.1. Objectives of the Experiment and Research Framework

The conducted experimental study is a pilot study aimed at experimentally verifying the proposed model and the developed software prototype for the automated generation and evaluation of test questions through generative artificial intelligence. The specific objectives of the experiment are:

1. To evaluate the ***ability of GenAI models to automatically generate correct and pedagogically adequate test questions***;
2. To analyze the ***ability of GenAI models to automatically evaluate the quality of test questions generated*** both by the same and by a different generative model;
3. To examine the ***influence of the selected language model on the quality of the generated questions*** and the reliability of their evaluation;
4. To demonstrate the ***applicability of the proposed model for multi-model generation and evaluation*** as a core component of the developed intelligent electronic testing platform.

The set of ***72 test questions*** used was not formed randomly, but was selected so as to cover different levels of difficulty and cognitive levels according to Bloom's taxonomy, thus providing a sufficient basis for the experimental verification of the proposed mechanisms for generation and evaluation.

In accordance with the objectives and scope of the pilot study, research questions were formulated, aimed at the experimental evaluation of the applicability of large language models in the automated generation and evaluation of test questions:

- ***RQ1***: To what extent can LLM models automatically generate correct, pedagogically adequate, and relevant test questions?
- ***RQ2***: To what extent can LLM models reliably evaluate the quality of automatically generated test questions?
- ***RQ3***: Are there differences in the performance of different LLM models in the generation and evaluation of test questions?
- ***RQ4***: How does the choice of the generation model (GenModel) and the evaluation model (EvalModel) affect the quality of the obtained results?

4.2. Experimental Design

The experiment includes four main stages:

- EP1: Generation of test questions by LLM models;
- EP2: Evaluation of the generated test questions by LLMs;
- EP3: Evaluation of the generated test questions by expert evaluators;
- EP4: Processing, analysis, and interpretation of the obtained results.

The first stage of the experiment is related to the ***generation of test questions by two selected LLM models: GPT-5.4 mini and Claude Sonnet 4.6***.

In the second stage of the experiment, ***the test questions are evaluated by LLM models acting as automated evaluators***. The models evaluate the same questions using an established evaluation rubric, without information about the origin of the questions.

The third stage aims to ***evaluate the generated test questions by expert evaluators***. The obtained expert evaluations are used as a reference standard against which the reliability and validity of the generated questions are assessed.

The fourth stage is focused on the ***processing, analysis, and interpretation of the data collected*** in the previous experimental phases.

The experiment was conducted within a single academic discipline in the field of computer science: "Databases".

Two large language models were used in the experiment: GPT-5.4 mini and Claude Sonnet 4.6. They performed clearly distinguished functional roles within the experimental design. Depending on the specific scenario, each model could participate as:

- ***GenModel*** – a model used for the automatic generation of test questions;
- ***EvalModel*** – a model used to evaluate the quality of test questions based on predefined criteria.

4.3. Procedures for Generating and Evaluating Test Questions

For the generation of test questions, multiple parameters were specified, reflecting real scenarios from teaching and assessment practice in higher education.

With each of the selected LLM models, an identical number of questions was generated: 18 questions per model, using the same prompt, automatically generated by the platform based on the parameters selected for the experiment.

Within this stage of the experiment, large language models were used in the role of automated evaluators. The main objective was to analyze their ability to evaluate the quality of test questions in a manner comparable to human expert evaluation. For each test question, the evaluation model (EvalModel) received the following information:

- *the text of the test question*, without information about the model that generated it, in order to prevent bias;
- *the correct answer*, provided as reference information necessary for the correct evaluation of the scientific accuracy and unambiguity of the question;
- *the evaluation rubric* used in the human expert evaluation, including the predefined 5 criteria and 10 indicators;
- *the five-point Likert scale* used for evaluation by the expert evaluators;
- *clear instructions not to rewrite, edit, or improve the question*, but only to perform evaluation according to the specified indicators.

An evaluation rubric was created to assess the quality of the questions generated by artificial intelligence. It contains 5 evaluation criteria, each of which includes 2 indicators (Table 1).

Table 1. Evaluation Rubric for Questions Generated by LLMs

Evaluation Criterion / Indicator	Weight
I. Reliability and Scientific Accuracy	
1. The question is scientifically accurate and does not contain factual errors.	15%
2. The correct answer is unambiguous and does not allow alternative interpretations.	10%
II. Alignment with the Learning Content	
3. The question is directly related to the provided learning content, such as text or a file.	10%
4. The question does not require knowledge beyond the scope of the provided learning material.	10%
III. Adequacy of the Difficulty Level	
5. The actual difficulty of the question corresponds to the specified level: easy, medium, difficult, or very difficult.	10%
6. The difficulty of the question derives from the content rather than from unclear wording.	10%
IV. Alignment with Bloom's Cognitive Level	
7. The question actually measures the specified cognitive level according to Bloom's taxonomy.	10%
8. The type of cognitive activity required by the question corresponds to the selected level, such as analysis or synthesis.	10%
V. Pedagogical Adequacy and Applicability	
9. The wording of the question is clear, unambiguous, and grammatically correct.	5%
10. The question is pedagogically adequate and can be used directly in a real test.	10%

The experimental design includes ***four scenarios*** (Table 2), which combine different large language models in the roles of test question generator (GenAI) and quality evaluator (EvalAI). The purpose of these scenarios is to examine both the individual behavior of the separate models and the effect of their interaction within the generation and evaluation process.

Scenario S1 combines GPT as the test question generator and Claude as the evaluator. In ***scenario S2***, the same LLM model, GPT, is used both for generating and evaluating the test questions. In ***scenario S3***, the roles are reversed, with Claude acting as the generator and GPT as the evaluator. ***Scenario S4*** includes Claude both as generator and evaluator.

Table 2. LLM–LLM Evaluation Scenarios

Scenario	**GenAI**	**EvalAI**
S1	GPT	Claude
S2	GPT	GPT
S3	Claude	GPT
S4	Claude	Claude

4.4. Analysis of the Results

The results from the evaluation of the generated questions across the examined scenarios are summarized and analyzed through a hierarchical approach, in which the scores are aggregated successively at three levels. First, scores are calculated for the individual indicators. Then, these scores are used to form criterion-level scores by applying the specified weights. Finally, the final weighted score for each scenario is derived.

Figure 12 presents the results for the scenarios in which a given LLM generates questions and the evaluation is performed by the same model, as self-evaluation, and by an expert. Figure 13 shows the results for the scenarios in which the questions are evaluated by another LLM and by an expert.

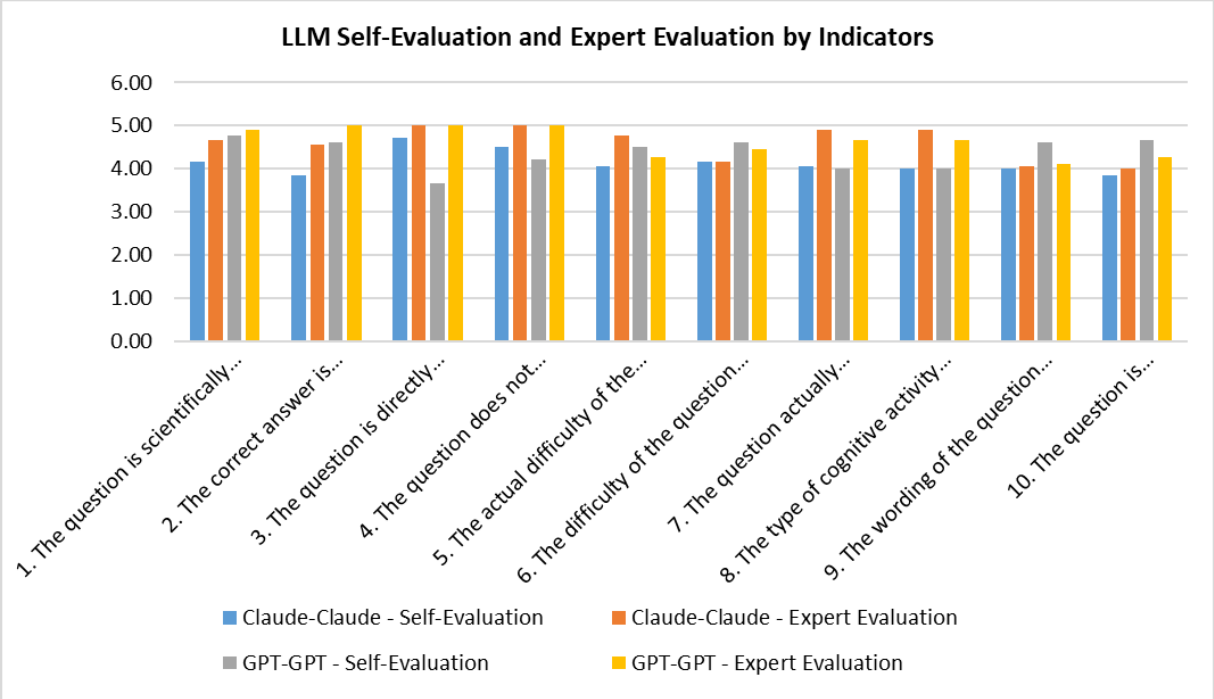

Figure 12. LLM Self-Evaluation and Expert Evaluation by Indicators

Overall, all average scores are above 3.0, and the majority of them fall within the range between 4.0 and 5.0. This indicates that, regardless of the scenario used, the generated questions are of high quality and largely satisfy the predefined criteria. In most cases, the expert evaluations are higher than the scores assigned by the LLM models. This shows that, from the perspective of the human expert, the generated questions are of acceptably high quality and are suitable for real educational use.

An interesting result is also observed in cross-evaluation. When GPT evaluates questions generated by Claude, corresponding to the Claude–GPT scenario, the scores are systematically lower than the expert evaluations. Conversely, when Claude evaluates questions generated by GPT, corresponding to the GPT–Claude scenario, the differences between the model's scores and the expert's scores are smaller.

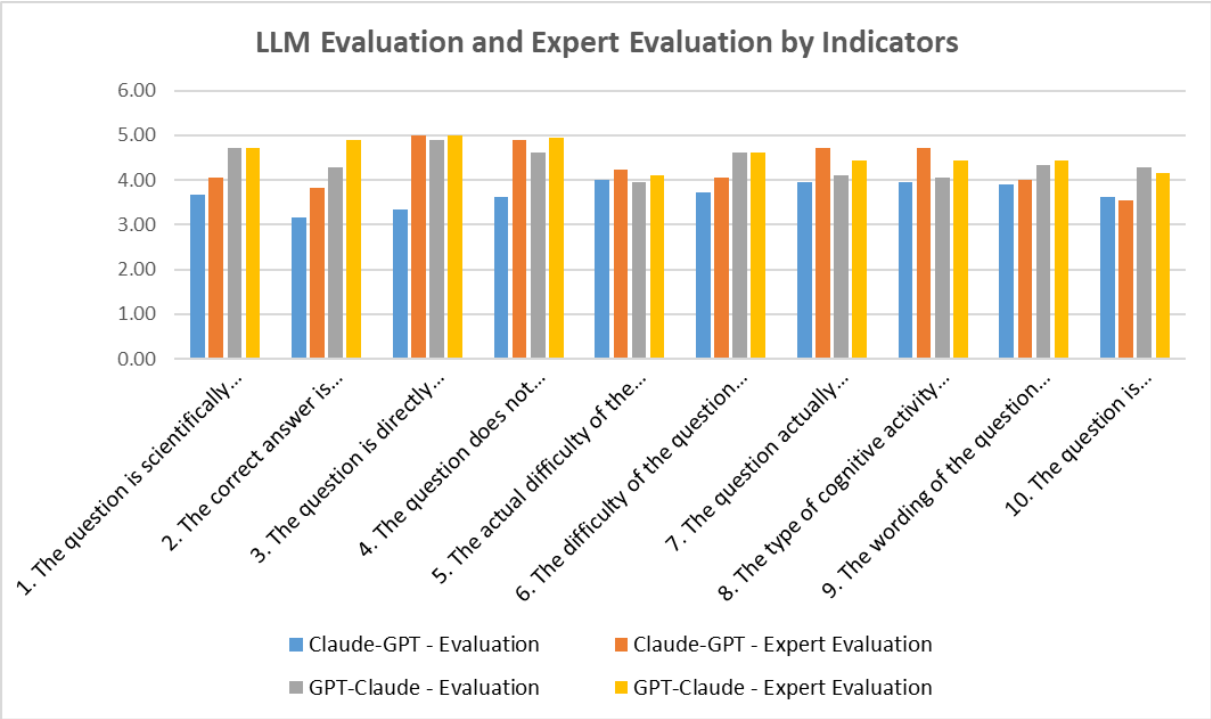

Figure 13. LLM Evaluation and Expert Evaluation by Indicators

For the criterion-level scores, the specified indicator weights, their corresponding scores, and normalization within the interval [1, 5] are used. The obtained results are presented in Fig. 14 and Fig. 15.

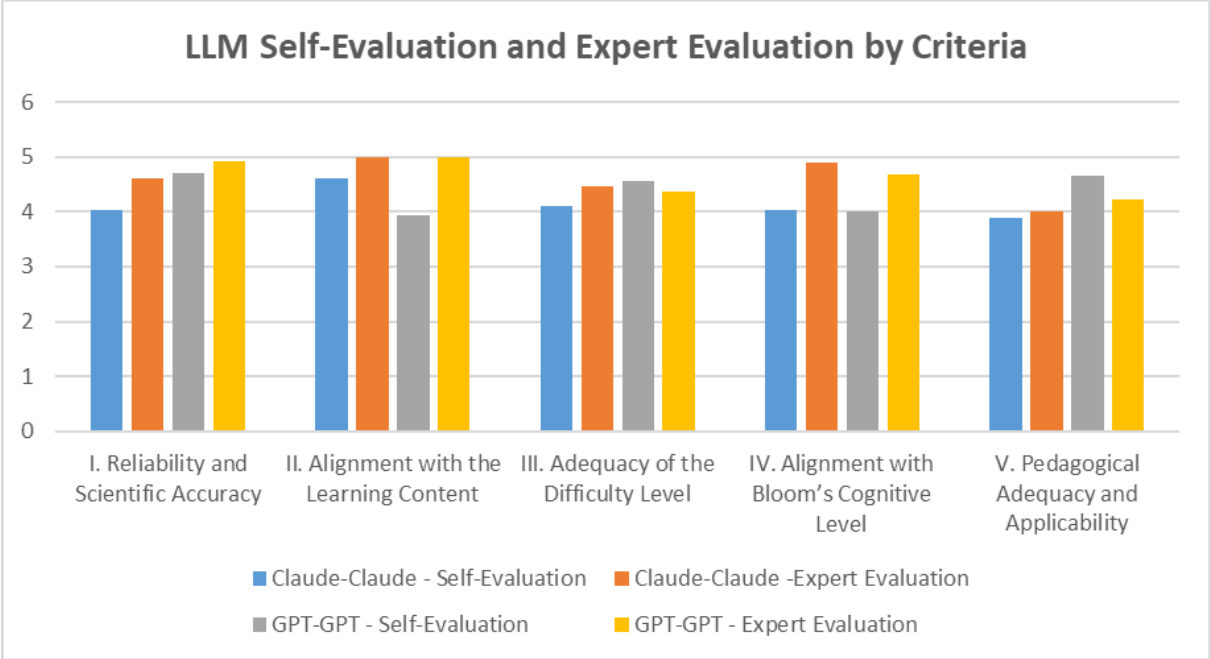

Figure 14. LLM Self-Evaluation and Expert Evaluation by Criteria

The analysis of the results by criteria confirms the observations made in the analysis of the individual indicators. The high scores across all scenarios show that the generated questions are of high quality. A tendency is also observed for the expert evaluations to be higher than the evaluations and self-evaluations provided by the models.

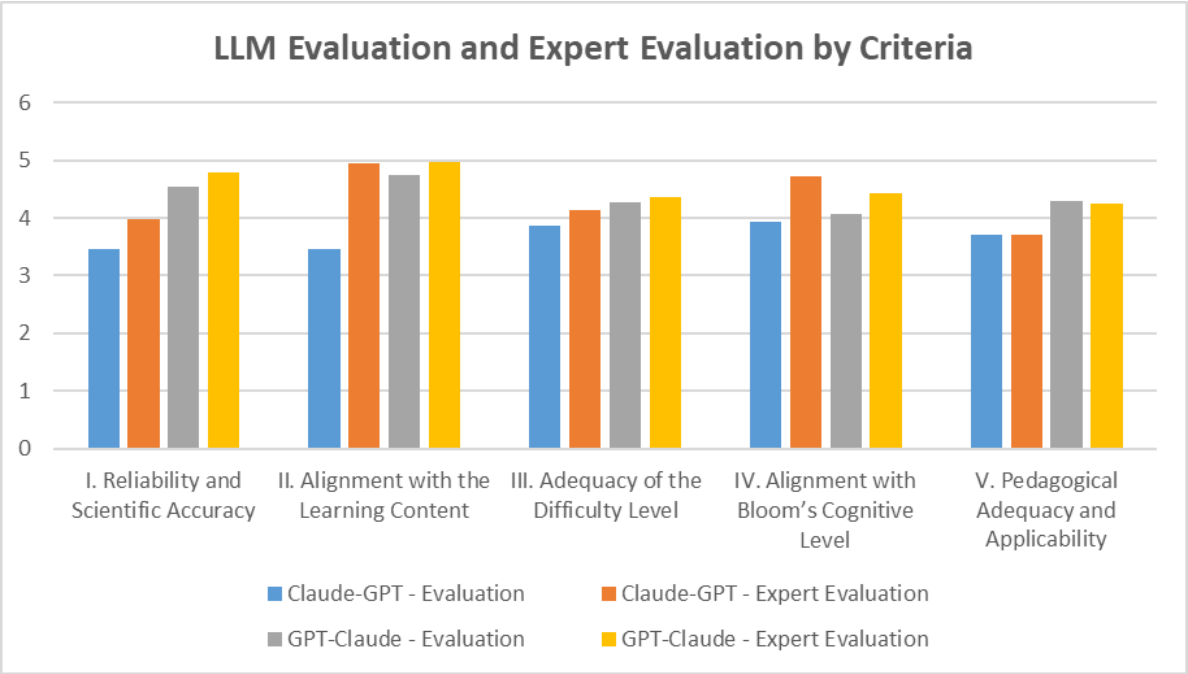

Figure 15. LLM Evaluation and Expert Evaluation by Criteria

The final weighted scores for all examined scenarios are presented in Fig. 16.

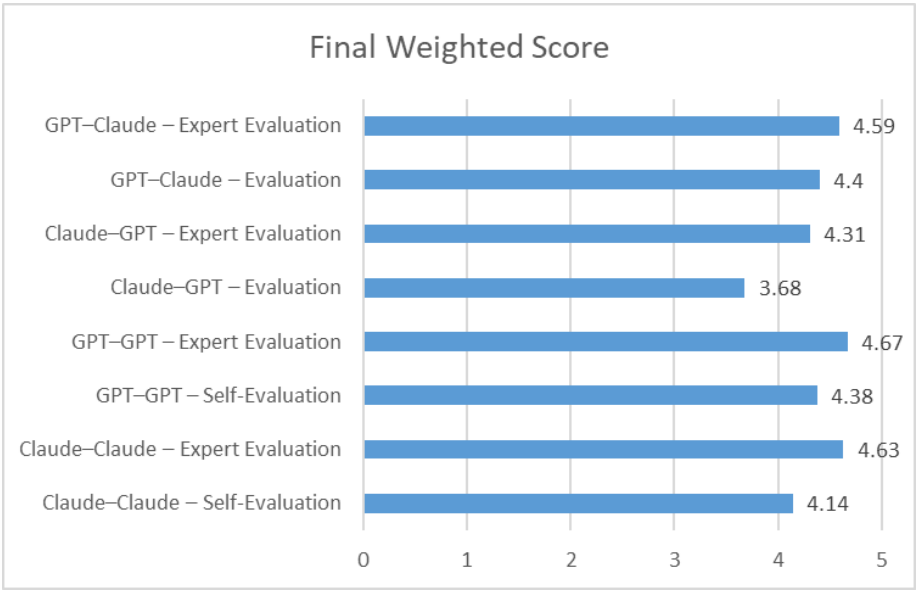

Figure 16. Final Weighted Score for All Scenarios

The obtained results show a high overall quality of the generated questions across all scenarios, with all expert evaluations exceeding 4.30.

Greater differences are observed in the self-evaluations and mutual evaluations performed by the models. The highest score assigned by an LLM was recorded in the GPT–Claude scenario (4.40), followed by GPT–GPT self-evaluation (4.38). The lowest result was obtained in the Claude–GPT scenario (3.68), while the Claude–Claude scenario achieved a score of 4.14.

Overall, the results show that both GPT and Claude are capable of generating high-quality questions, with a slight advantage for GPT.

4.5 Conclusions

The conducted experiments demonstrate the applicability and effectiveness of the proposed approach for using generative artificial intelligence in the creation, verification, and evaluation of test content. The obtained results confirm that large language models can successfully support the process of developing test questions while reducing the time and effort required from teachers.

CONCLUSION

The digital transformation of education and the rapid development of artificial intelligence create new opportunities for improving teaching and assessment processes. Modern educational environments require tools that not only automate routine activities but also support teachers in creating high-quality learning content and objectively assessing learners' knowledge. Generative artificial intelligence and large language models provide significant potential for the automated generation of test questions, analysis of results, and personalization of the assessment process. The present dissertation research is aimed at the development of an intelligent electronic testing platform that integrates the capabilities of generative artificial intelligence into a unified environment for creating, verifying, managing, and assessing test content.

Within the dissertation research, the following main tasks have been completed:

- Existing electronic testing systems have been studied, and the main functional and non-functional requirements for an intelligent electronic testing system have been identified;
- A conceptual model and architecture of an intelligent electronic testing platform have been developed, including mechanisms for the automated generation and verification of test questions through large language models;
- A module for automated adaptive assessment of learners has been developed, based on modern psychometric approaches and artificial intelligence;
- A software prototype of the proposed platform has been implemented;
- Experimental studies have been conducted to evaluate the quality and applicability of the developed solution.

Through the completion of the defined tasks, the main objective of the dissertation has been achieved: the development of a model, architecture, and software implementation of an intelligent electronic testing platform that uses generative artificial intelligence for the automated creation, verification, management, and assessment of test content.

Contributions

The main contributions of the dissertation can be classified as scientific, scientific-applied, and applied.

Scientific Contributions

- A conceptual framework of an intelligent electronic testing platform based on generative artificial intelligence has been created;
- An approach for multi-model verification of test questions through the use of independent AI models has been proposed.

Scientific-Applied Contributions

- A model of an intelligent platform for the automated generation of test questions through the use of generative language models and parameterized prompts has been developed;
- A mechanism for automated quality assessment of generated questions through independent AI verification has been proposed.

Applied Contributions

- A software prototype of an intelligent electronic testing platform has been implemented, including functionalities for the automated generation, verification, and management of test questions;
- Experimental studies have been conducted, proving the applicability of the developed solution.

Approbation

Participation in Research Projects

The results obtained during the research have been presented within three research projects:

- MUPD25-FMI-015, School for ICT Innovations: AI Technologies for the Transformation of Education (2025–2026).
- SP23-FMI-008, Development of Creative Potential in the School for ICT Innovations (2023–2024).
- MUPD23-FMI-021, Use of Artificial Intelligence Methods in Business (2023–2024).

With the financial support of the aforementioned projects, the results were reported at four scientific conferences.

Oral and Poster Presentations at Conferences

1. Ethical Challenges and Safety Strategies in the Use of Generative AI Systems in Education, Scientific Conference “Days of Science 2025”, November 21–22, 2025, Plovdiv.

2. The Future of Cloud-Based Learning: An AI-Based Architecture of a Software Application for Assessment, Scientific Conference “Days of Science 2023”, November 23–24, 2023, Plovdiv.

3. Algorithmic Generation of Test Questions from Text: An Overview of Contemporary AI Platforms and Their Application in the Educational Context, Scientific Conference “Days of Science 2023”, November 23–24, 2023, Plovdiv.

4. Artificial Intelligence Systems for Learning Management, XXXIII International Online Scientific Conference, June 1, 2023, Stara Zagora.

List of Publications Related to the Dissertation Topic

A total of six scientific publications have been published on the topic of the dissertation. Three of the publications are indexed in the internationally recognized databases Web of Science and/or Scopus. One publication is indexed in Web of Science with IF = 2.5 (Q2) and in Scopus with SJR = 0.52 (Q2):

1. Hadzhikoleva, S., T. Rachovski, E. Hadzhikolev, I. Ivanov, *The role of generative artificial intelligence in programmer training: opportunities and challenges*,

Proceedings of the 16th International Scientific and Practical Conference "Environment. Technology. Resources", June 19-20, 2025, Rezekne, Latvia. **(SCOPUS)** https://doi.org/10.17770/etr2025vol2.8608

2. Hadzhikoleva, S., T. Rachovski, I. Ivanov, E. Hadzhikolev, G. Dimitrov, *Automated Test Creation Using Large Language Models: A Practical Application*, Applied Sciences vol. 14, no. 19, 2024. **(Web of science Q2 IF=2,5; SCOPUS Q2 SJR=0.52)** https://doi.org/10.3390/app14199125
3. Rachovski, T., D. Petrova, I. Ivanov, *Automated Creation of Educational Questions: Analysis of Artificial Intelligence Technologies and Their Role in Education*, Proceedings of the 15th International Scientific and Practical Conference "Environment. Technology. Resources", June 27-28, 2024, Veliko Tarnovo, Bulgaria. **(SCOPUS)** https://doi.org/10.17770/etr2024vol2.8101
4. Ivanov, I., T. Rachovski, G. Pashev, *Integration of AI models for question generation and assessment: Application of ChatGPT and Gemini in education*, Scientific researches of the Union of Scientists in Bulgaria-Plovdiv, series B. Natural Sciences and the Humanities, Vol. XXV, ISSN 1311-9192 (Print), ISSN 2534-9376 (On-line), 2024, стр. 154-159. https://usb-plovdiv.org/wp-content/uploads/2024/12/2024_natural_sciences_and_humanities_vol_XXV.pdf
5. Ivanov, I., T. Rachovski, Algorithmic Generation of Test Questions from Text: An Overview of Contemporary AI Platforms and Their Application in the Educational Context, Scientific Works of the Union of Scientists in Bulgaria–Plovdiv, Series B. Engineering and Technologies, Vol. XXI, ISSN 1311-9419 (Print), ISSN 2534-9384 (Online), 2024, pp. 183–188. https://usb-plovdiv.org/wp-content/uploads/2024/05/2024_tehnika_i_tehnologii_tom_XXI.pdf
6. Ivanov, I., T. Rachovski, Artificial Intelligence Systems for Learning Management, Science and Technologies, Volume XIII, 2023, Number 2: Technical Studies, pp. 1–7. https://www.sustz.com/journal/2/2102.pdf

More than 30 citations of two of the publications related to the dissertation have been identified.

BIBLIOGRAPHY

(Agbo et al., 2020) Agbo, G.C.; Agbo, P.A. The role of computer vision in the development of knowledge-based systems for teaching and learning of English language education. *ACCENTS Trans. Image Process. Comput. Vis.* **2020**, *6*, 42–47. https://doi.org/10.19101/TIPCV.2020.618044.

(Baker & Kim, 2004) Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques*. Taylor & Francis Group.

, R.E.; Bejar, I.I. Validity and automad scoring: It's not only the scoring. *Educ. Meas. Issues Pract.* **1998**, *17*, 9–17.

(Biehl, 2016) Biehl, M. *RESTful API Design: Best Practices in API Design with REST*; ASIN: B01L6STMVW; API-University Press: 2016.

(Borisova, Hadzhikoleva et al., 2024) Borisova, M., S. Hadzhikoleva & E. Hadzhikolev. ChatGPT as a tool for creating educational games for school education in computer technology, International Conference on Virtual Learning, ISSN 2971-9291, ISSN-L 1844-8933, vol. 19, pp. 323-333, 2024. https://doi.org/10.58503/icvl-v19y202427

(Briganti & Le Moine, 2020) Briganti, G.; Le Moine, O. Artificial intelligence in medicine: Today and tomorrow. *Front. Med.* **2020**, *7*, 27. https://doi.org/10.3389/fmed.2020.00027.

(Bugbee, 1996) Bugbee, A.C. The Equivalence of Paper-and-Pencil and Computer-Based Testing. *J. Res. Comput. Educ*. **1996**, 28, 282–299. https://doi.org/10.1080/08886504.1996.10782166.
(Chaudhry, Sarwary et al., 2023) Chaudhry, I.; Sarwary, S.; Refae, G.; Chabchoub, H. Time to Revisit Existing Student's Performance Evaluation Approach in Higher Education Sector in a New Era of ChatGPT–A Case Study. *Cogent Educ.* **2023**, *10*, 2210461. https://doi.org/10.1080/2331186X.2023.2210461.
(Chen, Samuel et al., 2021) Chen, W.; Samuel, R.; Krishnamoorthy, S. Computer Vision for Dynamic Student Data Management in Higher Education Platform. *J. Mult. -Valued Log. Soft Comput.* **2021**, *36*, 5.
(Darayseh, 2023) Darayseh, A.A. Acceptance of artificial intelligence in teaching science: Science teachers' perspective. *Comput. Educ. Artif. Intell.* **2023**, *4*, 100132. https://doi.org/10.1016/j.caeai.2023.100132.
(Dempere, Modugu et al., 2023) Dempere, J.; Modugu, K.; Hesham, A.; Ramasamy, L. The impact of ChatGPT on higher education, *Front. Educ.* **2023**, *8*, 1206936. https://doi.org/10.3389/feduc.2023.1206936.
(Edmett, Ichaporia et al., 2023) Edmett, A.; Ichaporia, N.; Crompton, H.; Crichton, R. Artificial Intelligence and English Language Teaching: Preparing for the Future. British Council, 2023. Available online: https://www.teachingenglish.org.uk/sites/teacheng/files/2024-08/AI_and_ELT_Jul_2024.pdf
(Embretson & Reise, 2025) Embretson, S. E., & Reise, S. P. (2025). *Item response theory: Foundations for psychologists and social scientists*. Routledge. https://doi.org/10.4324/9781315726557
(Grassini, 2023) Grassini, S. Shaping the Future of Education: Exploring the Potential and Consequences of AI and ChatGPT in Educational Settings. *Educ. Sci.* **2023**, *13*, 692. https://doi.org/10.3390/educsci13070692.
(Hadzhikolev, Hadzhikoleva et al., 2021) Hadzhikolev, E.; Hadzhikoleva, S.; Yotov, K.; Borisova, M. Automated Assessment of Lower and Higher-Order Thinking Skills Using Artificial Intelligence Methods. *Commun. Comput. Inf. Sci.* **2021**, *1521*, 13–25.
(Hadzhikoleva, Borisova et al., 2025) Hadzhikoleva, S., Borisova, M., Hadzhikolev, E., & Raykova, Z. (2025, March). Building and Utilizing a Specialized Chatbot in Physics Education–Experiment and Results. In 2025 24th International Symposium INFOTEH-JAHORINA (INFOTEH) (pp. 1-6). IEEE. https://doi.org/10.1109/INFOTEH64129.2025.10959194
(Hadzhikoleva, Gorgorova et al., 2024) Hadzhikoleva, S., M. Gorgorova, E. Hadzhikolev, G. Pashev. (2024). AI-Driven Approach to Educational Game Creation, Proceedings of the 16th ICT Innovations Conference, 28-30 September 2024, Ohrid, Macedonia, ISBN 978-608-65468-4-7. https://ictinnovations.org/wp-content/uploads/2025/01/final_web_proceedings.pdf
(Hajkowicz, Sanderson et al., 2023) Hajkowicz, S.; Sanderson, C.; Karimi, S.; Bratanova, A.; Naughtin, C. Artificial intelligence adoption in the physical sciences, natural sciences, life sciences, social sciences and the arts and humanities: A bibliometric analysis of research publications from 1960–2021. *Technol. Soc.* **2023**, *74*, 102260. https://doi.org/10.1016/j.techsoc.2023.102260.
(Harry & Sayudin, 2023) Harry, A.; Sayudin, S. Role of AI in Education. *Interdiciplinary J. Hummanity* **2023**, *2*, 260–268. https://doi.org/10.58631/injurity.v2i3.52.
(Heilman, 2011) Heilman, M. Automatic Factual Question Generation from Text. Ph.D. Thesis, Carnegie Mellon University, Pittsburgh, PA, USA, 2011.
(Jeon & Lee, 2023) Jeon, J.; Lee, S. Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT. *Educ. Inf. Technol.* **2023**, *28*, 15873–15892. https://doi.org/10.1007/s10639-023-11834-1.
(Kandlhofer, Steinbauer et al., 2016) Kandlhofer, M.; Steinbauer, G.; Hirschmugl-Gaisch, S.; Huber, P. Artificial intelligence and computer science in education: From kindergarten to university. In Proceedings of the 2016 IEEE Frontiers in Education Conference (FIE), Erie, PA, USA, 12–15 October 2016. https://doi.org/10.1109/FIE.2016.7757570.
(Kasneci, Sessler et al., 2023) Kasneci, E.; Sessler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günnemann, S.; Hüllermeier, E.; et al. ChatGPT for good? On opportunities and challenges of large language models for education. *Learn. Individ. Differ.* **2023**, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274.

(Kucak, Juricic et al., 2018) Kucak, D.; Juricic, V.; Dambic, G. Machine Learning in Education–A Survey of Current Research Trends. In Proceedings of the 29th DAAAM International Symposium, Vienna, Austria, 24–27 October 2018. https://doi.org/10.2507/29th.daaam.proceedings.059.
(Mishra, 2019) Mishra, R. Usage of Data Analytics and Artificial Intelligence in Ensuring Quality Assurance at Higher Education Institutions. In Proceedings of the 2019 Amity International Conference on Artificial Intelligence (AICAI), Dubai, United Arab Emirates, 4–6 February 2019; pp. 1022–1025. https://doi.org/10.1109/AICAI.2019.8701392.
(Nunez & Lantada, 2020) Nunez, J.M.; Lantada, A.D. Artificial intelligence aided engineering education: State of the art, potentials and challenges. *Int. J. Eng. Educ.* **2020**, *36*, 1740–1751.
(Nurhayati & Halimah, 2024) Nurhayati, T.N.; Halimah, L. The Value and Technology: Maintaining Balance in Social Science Education in the Era of Artificial Intelligence. In Proceedings of the International Conference on Aplied Social Sciences in Education, Bangkok, Thailand, 14–16 November 2024; Volume 1, pp. 28–36. https://doi.org/10.31316/icasse.v1i1.6840.
(O'Connor, 2023) O'Connor, S. Open artificial intelligence platforms in nursing education: Tools for academic progress or abuse? *Nurse Educ. Pract.* **2023**, *66*, 103537. https://doi.org/10.1016/j.nepr.2022.103537.
(Osterlind, 1998) Osterlind, S.J. What is constructing test items? In *Constructing Test Items. Evaluation in Education and Human Services*; Springer: Dordrecht, The Netherlands, **1998**; Volume 25. https://doi.org/10.1007/978-94-009-1071-3_1.
(Pabitha, Mohana et al., 2014) Pabitha, P.; Mohana, M.; Suganthi, S.; Sivanandhini, B. Automatic Question Generation system. In Proceedings of the 2014 International Conference on Recent Trends in Information Technology, Chennai, India, 10–12 April 2014; pp. 1–5. https://doi.org/10.1109/ICRTIT.2014.6996216.
(Papasalouros, Kanaris et al., 2008) Papasalouros, A.; Kanaris, K.; Kotis, K. Automatic generation of multiple choice questions from domain ontologies. In Proceedings of the International Conference e-Learning 2008, Amsterdam, The Netherlands, 22–25 July 2008; pp. 427–434.
(Pradana, Elisa et al., 2023) Pradana, M.; Elisa, H.; Syarifuddin, S. Discussing ChatGPT in education: A literature review and bibliometric analysis. *Cogent Educ.* **2023**, *10*, 2243134. https://doi.org/10.1080/2331186X.2023.2243134.
(Rahiman & Kodikal, 2024) Rahiman, H.; Kodikal, R. Revolutionizing education: Artificial intelligence empowered learning in higher education. *Cogent Educ.* **2024**, *11*, 2293431. https://doi.org/10.1080/2331186X.2023.2293431.
(Rojas & Chiappe, 2024) Rojas, M.P.; Chiappe, A. Artificial Intelligence and Digital Ecosystems in Education: A Review. *Technol. Knowl. Learn.* **2024**, 1–18. https://doi.org/10.1007/s10758-024-09732-7.
(Sofianos, Stefanidakis et al., 2024) Sofianos, K.C.; Stefanidakis, M.; Kaponis, A.; Bukauskas, L. Assist of AI in a Smart Learning Environment. *IFIP Adv. Inf. Commun. Technol.* **2024**, *714*, 263–275. https://doi.org/10.1007/978-3-031-63223-5_20.
(Stokel-Walker, 2022) Stokel-Walker, C. AI bot ChatGPT writes smart essays-should academics worry? *Nature, 09 December 2022.* https://doi.org/10.1038/d41586-022-04397-7.
(Taecharungroj, 2023) Taecharungroj, V. "What Can ChatGPT Do?" Analyzing Early Reactions to the Innovative AI Chatbot on Twitter. *Big Data Cogn. Comput.* **2023**, *7*, 35. https://doi.org/10.3390/bdcc7010035.
(Van der Linden & Glas, 2010) Van der Linden, W. J., & Glas, C. A. (Eds.). (2010). *Elements of adaptive testing* (Vol. 10, pp. 978-0). New York: Springer. https://doi.org/10.1007/978-0-387-85461-8
(Wyse, 2010) Wyse, A. E. (2010). RJ DE AYALA (2009) The Theory and Practice of Item Response Theory. New York: The Guilford Press. ISBN: 978-1-59385-869-8. *Psychometrika*, *75*(4), 778-779. https://doi.org/10.1007/s11336-010-9179-z
(Xu & Ouyang, 2022) Xu, W.; Ouyang, F. The application of AI technologies in STEM education: A systematic review from 2011 to 2021. *Int. J. STEM Educ.* **2022**, *9*, 59. https://doi.org/10.1186/s40594-022-00377-5.
(Zhang & Aslan, 2021) Zhang, K.; Aslan, A.B. AI technologies for education: Recent research & future directions. *Comput. Educ. Artif. Intell.* **2021**, *2*, 100025. https://doi.org/10.1016/j.caeai.2021.100025.