



Transforming Assessment with Self-Correcting AI Pipelines: Towards Adaptive and Validated LMS Practices

Sergiy Yevseyev

Ivano-Frankivsk National Technical University of Oil and Gas, Ivano-Frankivsk, Ukraine

*sergiyevseyev@gmail.com

Abstract

The purpose of the study was to develop an effective approach to automated generation and validation of educational assessment materials in the context of distance learning using Large Language Models (LLM). While current research often overlooks the dynamic transformation of evaluation practices using sequential AI applications, this study addresses this gap by introducing a novel self-correcting AI pipeline that iteratively refines and validates questions based on structured feedback. As part of the study, an experimental methodology combined qualitative and quantitative methods, including computer modelling. Researchers conducted a comparative analysis of the results from three different LLMs: ChatGPT-4o, Google Gemini PRO 1.5, and Claude 3.5 Sonnet. Finally, the generated questions were validated based on pedagogical expediency, age compliance, and Bloom's taxonomy. To test the effectiveness of the approach, three stages of research were conducted. In the first stage, it was found that only 23% of questions generated without updated prompts met the stated criteria. In the second stage, after introducing clarifications to the prompts, this indicator increased to 63%. The highest results were achieved in the third stage, where an iterative hint refinement model using structured feedback was implemented: Claude 3.5 Sonnet achieved 92% of valid questions with the minimum number of clarifications (16), ChatGPT-4o 80% (40 clarifications), and Gemini 72%, but with the highest number of corrections (108). The process included question generation, answer verification, external validation, and iterative correction. The findings showed that effective AI-based development of multi-level computer science questions requires high-quality language models, clear instructions, and automated verification of cognitive level. The proposed algorithm enables educational institutions, platform developers, and teachers to generate assessment materials adapted to students' age and knowledge levels in distance or blended learning

Keywords: AI-driven assessment development; distance education; dynamic tracking; flexible verification; knowledge testing

How to Cite: S. Yevseyev, "Transforming Assessment with Self-Correcting AI Pipelines: Towards Adaptive and Validated LMS Practices", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 10, no. 4, pp. 1183–1196, Aug. 2026, doi: [10.29207/resti.v10i4.7191](https://doi.org/10.29207/resti.v10i4.7191).
Permalink/DOI: <https://doi.org/10.29207/resti.v10i4.7191>

Received: September 19, 2025

Accepted: May 23, 2026

Available Online: August 26, 2026

This is an open-access article under the [CC BY 4.0 License](https://creativecommons.org/licenses/by/4.0/)

Published by [Ikatan Ahli Informatika Indonesia](https://www.iaii.or.id/)

1. Introduction

In learning management systems (LMS), there is a deep restructuring of assessment approaches due to the introduction of sequential integrated artificial intelligence (AI) applications that use AI algorithms. Such tools transform conventional methods of knowledge testing by providing dynamic tracking of students' progress, personalized feedback, and flexible adaptation of tasks according to the individual learner's level of comprehension. This opens up new opportunities for improving the objectivity, transparency and efficiency of the evaluation process, which requires a systematic study of the transformations that occur under the influence of such technologies.

Recent literature establishes that integrating AI into Learning Management Systems (LMS) significantly enhances both assessment accuracy and personalized learning. While foundational studies demonstrate that AI-driven data analytics can successfully optimize learning trajectories and reduce the administrative burden on educators, these investigations predominantly treat AI as a static, one-way analytical tool [1], [2]. They often overlook the generative and interactive potential of Large Language Models (LLMs) to actively create, evaluate, and refine assessment content in real-time. Consequently, a critical gap remains in understanding how iterative, self-correcting AI pipelines can dynamically generate validated, multi-level cognitive tasks – a transformative

shift from static data management to active content generation that this study seeks to address.

Alotaibi [3] looked at the integration of AI and LMS in higher education, outlining opportunities and challenges for transforming educational processes, in particular, assessment practices, which have become more adaptive and flexible. He highlighted that the introduction of AI in LMS helps to solve problems of scalability of assessment and reduces subjectivity in decision-making by teachers. However, the researcher highlighted the difficulties associated with technical support and the ethical aspects of using AI in the educational environment. The researcher recommended developing strategies tailored to these challenges to ensure effective and secure AI integration. Amer-Yahia [4] explored AI-driven data approaches in educational environments, which contributed to a deeper analysis of learning outcomes and the development of an automatic feedback system. She pointed out the potential of AI to improve learning analytics, which helped to identify weaknesses in the knowledge of students.

Halkiopoulos and Gkintoni [5] conducted a systematic analysis of the use of cognitive neuropsychology in AI-supported e-learning platforms, finding that adaptive assessment, considering the cognitive characteristics of applicants, increased the accuracy and objectivity of assessments. They emphasised the importance of an interdisciplinary approach that combines AI and psychology to optimise learning processes. The recommendations included developing more sophisticated assessment models that take cognitive and emotional factors into consideration. Khrapatyi et al. [6], Huong [7] applied data analysis to assess the effectiveness of informatisation of vocational education, pointing out improvements in the productivity of LMS due to the introduction of AI applications. The researchers underscored that the use of Data Envelopment Analysis methods helped to objectively measure the results of introducing innovations in educational processes. The researchers pointed to the fact that informatisation in combination with AI contributed to improving the quality of knowledge control and monitoring the progress of applicants. They recommended further implementation of such technologies to improve the effectiveness of vocational education. Dandachi [8] explored the possibilities of personalised learning based on AI, with a focus on the sustainability of educational systems. He stressed that AI technologies allow adapting learning trajectories to the individual needs of students, which helps to increase the effectiveness of learning and maintain motivation in the long term. The researcher also considered the challenges of implementing such systems, in particular, issues of ethical use of data and ensuring equal access to technologies.

The impact of innovative AI technologies on distance learning of languages in universities was described by Sarnovska et al. [9], noting an improvement in the quality of evaluation due to automated tools. They also

highlighted the adaptability of the systems, which made it possible to adjust the curriculum in accordance with the individual characteristics of students. The researchers concluded that the introduction of AI has helped to reduce barriers to access to quality training during global challenges. The researchers stressed the need to further develop methodologies for integrating AI into language education. Semeniuk et al. [10] outlined new teaching methodologies using AI in Ukrainian higher education, which also addressed changes in assessment and feedback practices. They emphasised the role of AI in creating adaptive learning environments that meet current student requirements and needs. The researchers also noted the challenges associated with preparing teachers to work with new technologies. Vashishth et al. [11], and Jaiswal and Arun [12] investigated AI-driven learning analytics for personalised feedback and evaluation in higher education, emphasising the importance of algorithmic solutions to improve the quality of assessment and support students in the learning process. They pointed out that AI allowed them to quickly identify knowledge gaps and provide targeted recommendations. The results of the study confirmed the potential of AI in transforming conventional assessment practices into more dynamic and interactive models.

Despite the aspects highlighted by the above-mentioned researchers, the gaps included insufficient attention to the dynamics of transformation of evaluation practices in the context of using sequential applications based on AI. The impact of such technologies on the adaptability of assessment and the accuracy of tracking individual progress of students in LMS also remained limited. The purpose of the study was to investigate the possibilities of using AI to transform assessment practices in LMS. To achieve this goal, three tasks were set: to analyse the existing AI tools used to automate the creation of educational tasks; to evaluate the effectiveness of Large Language Models (LLM) in generating multi-level computer science questions; to investigate the impact of the introduction of adaptive and self-learning assessment systems on the quality of the educational process.

2. Methods

The study applied several methods, beginning with the generation of training tasks in the Moodle General Import Format Technology (GIFT) format. It also used an algorithmic separation for generation and the calculation of correct answers via an external calculator. Generated questions underwent automated validation for accuracy, logic, and compliance with Bloom's taxonomy. Additionally, tasks were structured in JavaScript Object Notation (JSON) format. Two independent language models semantically verified the correct answers, while three qualified specialists provided expert evaluation if significant discrepancies arose. In case of significant discrepancies between the validating models, the question was submitted for

manual expert review. The review was conducted by three qualified specialists who assessed pedagogical expediency, age appropriateness, and conceptual correctness. The above methods were aimed at creating and validating assessment materials for distance learning using LLM. At the first stage, general system architecture was developed, which allowed dividing the generation process into three main stages: the initial formulation of the task, external verification of the generated response, and iterative improvement of the initial request. Then three modern LLM were selected – ChatGPT-4o, Google Gemini PRO 1.5, and Claude 3.5 Sonnet – for comparative use in experiments. These tools were chosen because of their relevance, high performance and recognition as one of the most powerful modern models of AI in the field of text generation and information processing. The comparison was based on the criteria of response accuracy, logical reasoning ability, content relevance, language quality, and request processing speed. The transformation of assessment practices involved not only automating the task creation process using AI, but also shifting the assessment paradigm: from conventional static questions to dynamic and adaptive ones that adjust to the learner's level.

At the second stage, questions of varying complexity were generated: from simple algebra problems designed in the Moodle GIFT format, to complex computer science test queries focused on high cognitive levels according to the Bloom's Revised Taxonomy (BRT), in particular, analysis, application, and evaluation. The task formulation was standardised to meet pedagogical goals and cognitive complexity requirements in accordance with BRT. For example, a generalised instruction template looked like this: “create a test question on a subject (algebra or computer science) for high school students”. The response format is (GIFT or JSON). The question should be formulated in English, use language appropriate for the specific age group, contain (one or more) correct answers and correspond to the cognitive level of “application”, “analysis”, or “assessment”. This approach ensured the unity of experimental conditions and reproducibility of the results. Further, the generated materials were validated by automated verification of accuracy, clarity, and compliance with pedagogical goals. The review covered compliance of the wording with the educational level, logical structure, language clarity, and compliance with BRT. If errors or inconsistencies were detected during verification, the algorithm automatically optimised the initial request based on the feedback received. This iterative process of reformulating questions through repeated API calls (Application Programming Interface) ensured error reduction, increased content relevance, and compliance with the level of training of the target audience. All stages of the study were implemented in compliance with the principles of openness and reproducibility, and using a cloud environment to automate the operation of models, which reduced the need for manual

intervention and simultaneously maintained quality control. To visualise the methodological sequence, the workflow of AI-based generation and validation of assessment materials is presented in Figure 1.

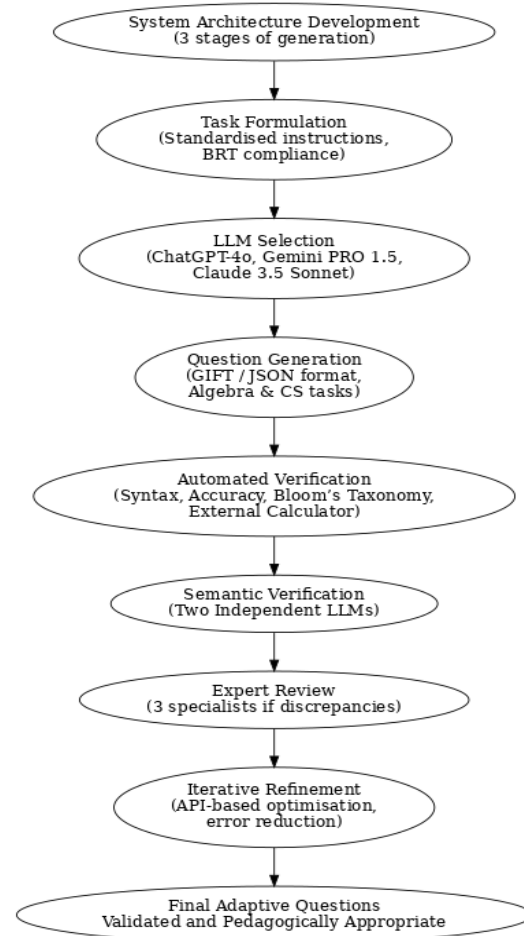


Figure 1. Workflow of AI-based generation and validation of assessment materials

The workflow progresses sequentially, beginning with the System Architecture Development, which encompasses the three stages of generation. This is followed by Task Formulation, which relies on standardised instructions to ensure compliance with Bloom's Revised Taxonomy (BRT). Next, LLM Selection takes place, specifically identifying ChatGPT-4o, Google Gemini PRO 1.5, and Claude 3.5 Sonnet as the primary tools. The selected models then proceed to Question Generation, producing algebra and computer science tasks structured in either GIFT or JSON formats. Following generation, the materials undergo Automated Verification to check syntax, mathematical accuracy via an external calculator, and continued alignment with Bloom's Taxonomy. The next step is Semantic Verification, which is conducted by two independent language models. If significant discrepancies arise during this verification, the process triggers a manual Expert Review evaluated by three qualified specialists. The workflow then moves into Iterative Refinement, employing API-based optimisation to significantly reduce errors based on the

feedback received. Ultimately, this comprehensive sequence results in Final Adaptive Questions that are fully validated and pedagogically appropriate.

Experiment 1. Generate questions with one correct answer in Moodle GIFT format. In the first experiment, the LLM algorithm assumed the following task: generate 100 algebraic problems in GIFT format. To do this, each of the models (ChatGPT-4o, Claude 3.5 Sonnet, and Gemini PRO 1.5) was sent requests with instructions to generate questions with one correct option among several suggested ones. The requests included clear wording about the format and topic – elementary arithmetic operations at the middle school level. The instructions set the task to generate 100 algebra problems in GIFT format with one correct answer among several suggested ones, the topics were elementary arithmetic operations and first-degree equations at the middle school level.

After generating the problems, each question was checked according to three criteria: compliance with the GIFT format syntax, mathematical accuracy (confirmed by a third-party calculator), and correct labelling of options (so that only one answer is marked as correct). Detected errors were automatically recorded and categorised (for example, the absence of a correct answer, the presence of several correct ones, or format failures).

3. Results and Discussions

To explore the potential of large language models (LLMs) in generating high-quality assessment materials for distance learning, a series of experiments was conducted. Each subsequent stage built on the findings and limitations of the previous one, gradually shifting the focus from syntactic correctness and error reduction toward conceptual depth, pedagogical value, and validation mechanisms:

Experiment 1 (updated). Separating problem generation and calculating the correct answer. In an improved version of the first experiment, algorithmic role separation was introduced: LLMs were responsible only for generating the problem text, while the calculation of the correct answer was performed by an external calculator. An external calculator was used to ensure the objectivity and accuracy of calculations, since LMS was responsible exclusively for generating the text of the problem, and not for solving it. The script automatically calculated the correct answer and added it to the question template as the only correct option. Additional false variants were created either by the model itself, or pseudorandom based on mathematical deviations. All variants were ordered in the GIFT format, considering the syntax rules. This approach significantly reduced the number of mathematical and format errors.

Experiment 2. Questions with multiple correct answers, no GIFT, with validation by other LLM. In the second experiment, the algorithm was modified to meet the

new goals. The models were assigned to generate questions from the computer science field for high school, where there would be several correct answers. The GIFT format was no longer used; instead, structured JSON was used, which included: the question itself, a list of correct answers, and a list of incorrect ones. The models did not receive hints about which options should be correct – this was tested separately. After generation, the question was automatically validated using two independent language models. These models rated each answer as correct or incorrect based on the semantics, and the general meaning of the question: compliance with academic level, thematic accuracy, and cognitive level (according to BRT). If the testing models had significant differences, the question was submitted for expert review. Three highly qualified experts with substantial experience in educational technologies conducted this review. Each expert was a candidate of Pedagogical Sciences with at least 10 years of school teaching experience. They also had practical experience implementing digital tools, developing educational programmes, and executing innovative educational projects nationally and internationally.

This algorithm allowed to shift the focus to the conceptual quality and depth of questions, and not just to the syntax. The questions were increasingly answered with the levels of “analysis”, “evaluation” or “application”. Despite the lack of a rigid template, LLM coped well with setting pedagogically valuable tasks, although sometimes there were problems with a clear distinction between correct and incorrect options. To address these issues, Experiment 2 simplified the task by replacing the GIFT format with a requirement for the language model to generate a Computer Science/IT question for high school students, providing multiple correct and incorrect answers in a structured JSON format. Creating high-quality multiple-choice science questions goes beyond simple content accuracy [13], [14]. The BRT provides a solid basis for classifying the cognitive depth of questions created. BRT describes a hierarchy of cognitive processes: memorisation, comprehension, application, analysis, evaluation, and creation. For example, the task “List all programming languages” corresponds to the “memorisation” level, while the question “How will system performance change after implementing a new algorithm?” refers to the “analysis” or “evaluation” levels.

Experiment 3. Interactive query improvement and self-correction based on feedback. In the third experiment, a more complex, multi-step algorithm was implemented, which provided for iterative generation and step-by-step improvement of queries. Initially, the model created a list of educational topics in accordance with the Computer Science curriculum in school. Then, for each topic, it separately generated questions, a list of correct answers, and false options – all in JSON format. The JSON structure used in the experiment includes the following key elements:

1. “Question” – contains the text of the question posed to the learner (e.g., “Which of the following are components of a computer system?”).

2. “Type” – specifies the question format, such as “multiple_select” (indicating that multiple correct answers are allowed).

3. “Options” – an array of answer choices, where each option includes:

– “text” – the content of the answer option (e.g., “CPU”, “Printer”);

– “correct” – a boolean value (true or false) indicating whether the option is correct.

4. “Metadata” – a block of contextual information about the question, including:

– “subject” – the academic subject (e.g., “Computer Science”);

– “grade_level” – the intended education level (e.g., “Middle School”);

– “cognitive_level” – the cognitive complexity level according to Bloom’s Taxonomy (e.g., “understanding”, “analysis”);

– “validation” – verification information, containing;

– “verified_by” – a list of language models used to automatically verify the question (e.g., “ChatGPT-4o”, “Claude 3.5 Sonnet”);

– “verification_status” – the current status of the verification process (“pending”, “verified”, or “rejected”).

This structure is particularly useful for generating educational content that is both human- and machine-readable. It can be directly integrated into LMS for formative assessment, automated quiz creation, and adaptive testing. Additionally, the metadata and validation components allow for quality control via AI-driven peer-review, ensuring the reliability of generated questions before they are presented to students [15]. This makes the structure suitable not only for AI-assisted question generation, but also for scalable deployment in real-world digital education platforms.

At each stage, the result was checked by other LLM or manually. When the model made a mistake (for example, an insufficiently clearly formulated question or an incorrect answer), it was provided with structured feedback in the form of a comment or a list of problems. The LLM then had to refine the initial prompt itself, taking these comments into consideration, and regenerate the questions. This is how a self-learning cycle was formed, in which each new request was more accurate than the previous one. Three models were used in the experiment: ChatGPT-4o, Google Gemini 1.5, and Claude 3.5 Sonnet. Each of them generated 100 questions, after which the questions were tested for compliance with restrictions (compliance with the topic, lexical simplicity, taxonomic level, availability of

clearly defined correct and incorrect answers). In case of errors, the models received structured feedback, for example: “The term “recursion” requires additional explanation for high school students. The answer contains a mention of “tail-call optimization”, which is too complex a concept for this age category”.

In addition to quality, the volume of API requests was also evaluated. Each full cycle requires at least three requests to the base model (questions, correct, incorrect options) and up to 10 requests to the validation models. If there were errors, the cycle was repeated up to 5 times, which created a large load on computing resources. Thus, for a scalable implementation of an adaptive evaluation system, it is advisable to consider using local models or optimising the number of requests.

The frequency of typical errors and question generation problems in the table was determined by a quantitative analysis of the experiment's results. First, each model generated a set of questions, which were then checked for specific errors. Finally, the number of errors of a certain type was divided by the total number of generated questions to express the result as a percentage. The study identified an algorithm for using AI to create adaptive, pedagogically sound assessment systems in the context of distance learning.

Practical implementation recommendations were formulated based on the obtained results: an algorithm was proposed for integrating AI into the creation of adaptive, pedagogically sound assessment systems in the context of distance learning, taking into account the identified typical errors in question generation and their frequency.

In a conventional LMS, assessment was usually performed using standard tests created manually. This required considerable effort from teachers, especially those who taught in multiple classes, in different courses or faculties. In the process of transformation of assessment systems in the field of education, modern learning management platforms play a key role, significantly expanding their functionality through the integration of intelligent tools based on AI [16-18]. These platforms automate numerous stages of the evaluation process – from generating test tasks to verifying them and providing detailed feedback, which improves the quality of evaluation and reduces the load on system users. Automated modules in LMS can adapt tasks to different levels of complexity. They can also provide iterative clarification of questions based on an analysis of the given answers. Furthermore, these modules perform in-depth analyses of learning outcomes, which allows educators to track individual progress and identify gaps in knowledge. In addition, advanced LMS support recognition of various response formats, text, numeric, handwritten, and multimedia, which expands the possibilities for evaluating various competencies. Integration of AI applications allows not only automating routine operations, but also to

significantly increasing the level of personalisation of the educational process, offering users individual recommendations for self-learning and development [19], [20]. The platforms are able to generate dynamic evaluation scenarios that respond to current results, which contributes to a more flexible and adaptive approach to learning. LMS analytical tools provide detailed feedback that goes beyond simply evaluating the correctness or incorrectness of an answer, offering information about common mistakes, the level of assimilation of the material, and recommendations for further development of specific topics [21], [22].

The functionality of modern LMS also includes integration with various external services and APIs, which allows expanding the range of available assessment and training tools, including generating questions based on LLM, automatic scanning of handwritten works, and conducting interactive surveys and quizzes with instant analysis of results. Such systems support scalability and flexibility, which is especially important for distance and mixed learning, ensuring the continuity of the educational process even in difficult conditions. In general, advanced LMS with built-in AI tools transform conventional assessment practices, making them more efficient, operational, and focused on individual user needs, which significantly improves the overall quality and accessibility of education [23], [24]. This creates a foundation for further research and innovation in the field of automated creation and validation of educational materials, considering the pedagogical requirements and cognitive levels of participants in the educational process.

Transformation of assessment practices in Ukraine using AI-based tools not only reduces the burden on teachers, but also improves the quality of feedback, evaluation accuracy, and overall effectiveness of distance learning. This corresponds to the new realities of Ukrainian education and creates the foundation for further modernisation of the system. In Table 1, a gradual assessment of the LLM's ability to generate educational content allows tracking the dynamics of the quality of tasks created according to certain criteria.

Table 1 shows the comparative results of the gradual assessment of the abilities of three modern LLM – ChatGPT-4o, Gemini PRO 1.5, and Claude 3.5 Sonnet – to generate educational content of varying complexity, including simple examples, test questions, analytical tasks, and complete sets in GIFT format. The results showed the overall advantage of ChatGPT-4o, which demonstrated the highest accuracy (up to 54% of correct questions), stable generation of formatted content even at difficult levels, and the lowest average response time (1.94 seconds). This advantage stems from ChatGPT-4o's ability to maintain logical clarity and strict formatting accuracy when processing structured prompts. In contrast, Gemini PRO 1.5 and

Claude 3.5 Sonnet were inferior in both accuracy and structure in this initial phase, especially at higher cognitive levels of tasks. These limitations are largely caused by their sensitivity to rigid formatting constraints; without highly specific structural limitations embedded in the initial prompt, these models struggle to adhere to strict syntax requirements, frequently resulting in format violations and an inability to properly separate correct responses from distractors. Table 2 shows an improvement in the accuracy of training content generation after configuring queries and considering format constraints. This increase in accuracy can be formally expressed and calculated by comparing the results with Table 1 using the following formula:

$$\text{Improvement (\%)} = \left(\frac{A_{\text{updated}} - A_{\text{initial}}}{A_{\text{initial}}} \right) \times 100, \quad (1)$$

A_{updated} is the percentage of correct questions in the updated experiment (Table 2), and A_{initial} is the baseline percentage from the first experiment (Table 1). This calculated improvement is especially noticeable in improving the quality of output data in the ChatGPT-4o and Claude 3.5 Sonnet models.

Table 2 compared three advanced language models by three key parameters: API error rate, percentage of correctly generated questions, and average runtime. The average time value was obtained by logging the duration of each API request, from the moment the prompt was dispatched to the model until the complete generated response was received, and calculating the arithmetic mean of these recorded durations across all 100 generated questions in the experimental set. The Claude 3.5 Sonnet model showed the highest accuracy – 100% correct questions, but spent more time generating than ChatGPT. ChatGPT-4o ran the fastest (averaging 4.35 seconds per request) and showed a high level of accuracy (81%), although its API error rate was slightly higher (5%). Gemini PRO 1.5 had the lowest percentage of correct answers (47%) and the longest average generation time (11.45 seconds), despite the same level of API errors as ChatGPT. These results showed a balance between the speed, accuracy, and stability of each model.

During experiments to test the ability of LLM to generate questions in the Moodle GIFT format, it was found that models often made mistakes, in particular, in the formatting of the answer and the GIFT structure. These errors caused the need for repeated requests to the API, which significantly increased the time and cost of calculations. Analysis of the frequency of such errors helped to determine how stable the work of models is in real-world application conditions, and to outline areas for further improvement of generation algorithms to reduce the number of repeated calls and improve the quality of automatically generated educational content.

Table 1. Gradual Assessment of the LLM's Ability to Generate Educational Content

Level	Task type	Model	Sample task	Example of the result (abbreviated)	Comment	API error rate	Correct questions (%)	Average time (s)
1	Generating simple examples	ChatGPT-4o	Solve: 5x-10=15	::Math::What is x if 5x-10=15? {=5}	Correct solution in GIFT format	4%	54%	1.94
		Gemini PRO 1.5	Solve: 5x-10=15	::Math::What is x if 5x-10=15? {=5}	Required format clarification	3%	27%	4.1
		Claude 3.5 Sonnet	Solve: 5x-10=15	Simple explanation, no formatting GIFT	Format not followed	2%	31%	3.3
		Claude 3.5 Sonnet	Which of the following is a function of parliament?	::Gov::What is a function of parliament? {~Executive=Legislation ~Judicial}	Semantic inaccuracies are sometimes present	2%	31%	3.3
2	Test questions with a choice	Gemini PRO 1.5	Which of the following is a function of parliament?	::Gov::What is a function of parliament? {=Legislation ~Taxation ~Repression}	Answers are sometimes confusing	3%	27%	4.1
		ChatGPT-4o	Which of the following is a function of parliament?	Correct generation with clear separation of responses	The most balanced result	4%	54%	1.94
		ChatGPT-4o	Compare the benefits of online and offline training	Generated task with evaluation criteria	High level of cognitive construction	4%	54%	1.94
		Gemini PRO 1.5	Evaluate the impact of AI on students' independent work	Task is correct, but without criteria	Requires revision	3%	27%	4.1
3	Generation of analytical tasks	Claude 3.5 Sonnet	Analyse the role of media in the educational process	Generating text without an evaluation approach	Less structured wording	2%	31%	3.3
		ChatGPT-4o	Generate 5 mixed questions (arithmetic, logic...)	GIFT file with correct formatting of all 5 questions	Highest level of compliance	4%	54%	1.94
		Gemini PRO 1.5	Same task	4 out of 5 questions in the format, one with an error	Minor structural errors	3%	27%	4.1
		Claude 3.5 Sonnet	Same task	3 out of 5 questions in the format, the rest – free text	Least stable response in a complex task	2%	31%	3.3
4	Generating a full GIFT set							

Table 2. Results of Experiment 1 (updated)

Model	API error rate	Correct questions	Average time (s)
ChatGPT-4o	5%	81%	4.35
Google Gemini PRO 1.5	5%	47%	11.45
Claude 3.5 Sonnet	2%	100%	8.32

After the results of the first experiment (in a modified form), it became obvious that both ChatGPT-4o and Google Gemini PRO 1.5 had difficulty generating the correct questions in the Moodle GIFT format. Format errors, such as incorrect character placement and multiple pseudo-correct response options, have demonstrated that these models face significant difficulties in meeting strict syntax requirements.

Modern language models are familiar with BRT, and taxonomy levels are actively used by researchers in various fields [25], [26]. In the field of Computer Science/IT, questions at the level of “memorisation” and “understanding” covered basic concepts: computer components or stages of the software development lifecycle. At higher levels of the hierarchy, questions at the “application” or “evaluation” level required students to use their knowledge in new situations: debugging code or configuring the network to meet certain requirements. Questions at the “analysis” level encouraged students to explore relationships in complex systems: for example, compare the

performance of different sorting algorithms or database management systems. The highest levels (“evaluation/creation”) required students to make decisions or create a unique product based on knowledge: analysing security aspects of the architecture or modelling a new software solution. This approach is useful for teachers: by learning the cognitive depth of AI-generated questions, they can choose the most appropriate ones for their audience. In particular, Table 3 shows the results of evaluating the ability of LLM to generate computer science and it questions in JSON format in the second experiment.

Table 3. Evaluation of the LLM’s Ability to Generate Computer Science/It Questions in JSON Format (Experiment 2)

Model	Sample question	JSON structure	BRT level	Quality of wording	Comment
ChatGPT-4o	“Which of these options is the advantage of SSD over HDD?”	+	Understanding	High	Clear structure, logical question, correct classification
Claude 3.5 Sonnet	“Assess the risks associated with using open Wi-Fi networks”	+	Evaluation	High	Advanced level of thinking, high-quality JSON structure
Gemini PRO 1.5	“Select the correct statements about operating system types”	Partially	Memorisation	Average	Some JSON elements were incorrect, and the responses were blurry
ChatGPT-4o	“What happens if the code changes the variable type from int to float?”	+	Analysis	High	Successful statement, relevant task for practice
Claude 3.5 Sonnet	“Compare the performance of two sorting algorithms on large data sets”	+	Analysis	High	A deep task, an example of critical thinking
Gemini PRO 1.5	“Specify all peripherals that transmit information to the computer”	-	Memorisation	Low	The JSON structure is incorrect, and some of the options are incorrect
Claude 3.5 Sonnet	“Create a simple example of a phishing attack in an email”	+	Creation	High	Very high-quality task at the highest cognitive level

Objectives:

Prompt Refinement by LLM: To evaluate LLM’s ability to improve the prompt, producing better results.

Accuracy of Generated Questions: To assess whether LLM is capable of generating questions using a refined prompt, that faithfully represent the specifications.

The evaluation results of the LLMs’ ability to generate Computer Science and IT questions in JSON format, presented in Table 3, indicate significant differences in the quality and complexity of the generated questions. ChatGPT-4o demonstrated high-quality wording and the ability to generate clear, logically structured questions covering various cognitive levels, from

“understanding” to “analysis”, with examples of relevant and practical tasks Claude 3.5 Sonnet also showed a high level by creating deep and critical questions, with a particular focus on evaluation and creation that meets high cognitive requirements. Gemini PRO 1.5 showed more limited features, with noticeable shortcomings in the correctness of the JSON structure and ambiguity of answers, which reduced the overall quality of questions, especially at memory levels. This indicates greater stability and adaptability of the first two models in creating multi-level and meaningful tasks necessary for effective educational assessment. Table 4 shows the effectiveness of iterative refinement and validation of questions in LMS.

Table 4. Effectiveness of Iterative Clarifications and Validation of Questions

Model	Clarification of questions	Clarification of the correct ones	Clarification of distractors	Failures (%)	Valid questions (%)
ChatGPT-4o	40	28	12	20%	80%
Google Gemini 1.5	108	8	36	28%	72%
Claude 3.5 Sonnet	16	0	16	8%	92%

The results presented in Table 4 demonstrate different strategies and performance levels for iterative refinement and validation of questions by three LLM. ChatGPT-4o showed a balanced approach, with 40

clarifications in total, where 28 concerned correct answers and only 12 – distractors; with a failure rate of 20%, and the share of valid questions reached 80%, which indicates moderate stability with high content

quality. On the other hand, Google Gemini 1.5 showed a more aggressive refinement strategy (108 cases), but with a significant bias towards distractors (36 against only 8 clarifications of correct answers), which obviously reduced the quality – with 28% failures, only 72% of questions were found to be valid. This high failure rate occurs because Gemini struggles to natively interpret semantic nuances, leading it to generate ambiguous answers that require extensive external intervention to restructure JSON elements and clarify distractors. The highest rate of valid questions (92%) was provided by the Claude 3.5 Sonnet model, which showed the lowest number of clarifications (16 overall, all focused on distractors) and the lowest failure rate (8%). Claude's high validity rate is driven by its inherent capacity to internalize higher-order cognitive instructions, such as analysis and creation according to Bloom's Taxonomy, allowing it to construct deep, logically sound tasks on the first attempt without triggering cyclical error correction loops. Table 5 shows typical errors and problems that occurred during the generation of questions at each stage of the process.

Table 5. Typical Errors and Problems of Generating Questions at Each Stage

Error type/problem	Gemini 1.5	ChatGPT-4o	Claude 3.5 Sonnet
Generating a question instead of an updated hint	65	15	5
Failure to comply with all restrictions	70	25	7
Complexity of topics / non-standard topics	60	27	20
Incorrect separation of right and wrong	25	35	10

Table 5 illustrates exactly what difficulties language models most often encountered during iterative generation. Gemini 1.5 often mistakenly replaced the updated hint with a new question; this complication is a direct consequence of the model losing the distinction between generating original content and executing iterative feedback when specific structural constraints are lacking. Claude 3.5 Sonnet had the fewest errors of all types, which partly explains its highest success rate, as its architecture is better suited to maintaining the semantic boundaries required for complex evaluation tasks and advanced JSON structures. All models sometimes encountered problems in determining the complexity of topics or confused correct and incorrect answers. This indicates a broader algorithmic limitation: when models attempt to merge multiple cognitive levels within a single task, they often fail to maintain clear didactic distinctions, underscoring the necessity of integrated automated validators to flag these issues and prevent formatting decay. Experiment 3 confirmed that iterative improvement of hints significantly improves the quality of question generation in LLM. Claude 3.5 Sonnet proved to be the most effective, while Google Gemini, despite its lower accuracy, demonstrated flexibility and adaptability due to more refinements. ChatGPT-4o remained a stable

and practical option with a moderate level of intervention.

The study identified an effective algorithm for using AI to create adaptive, pedagogically sound assessment systems, which is especially relevant in the context of distance learning. In particular, based on three experiments, a step-by-step methodology has been developed that allows LLM not only to generate learning questions, but also to independently correct and improve them in accordance with clearly defined didactic requirements. The main focus was on creating computer science test questions for high school students, considering Bloom's taxonomy, material complexity, pedagogical feasibility, and students' cognitive level.

In Experiment 1, the basic ability of LLM to create questions without specifying hints was tested. It turned out that the vast majority of the generated questions either had unclear wording, or did not correspond to the age characteristics of students. In particular, in 58% of cases, the questions contained linguistic or conceptual errors. In addition, only 23% met the criteria for cognitive levels of "application", "analysis", or "evaluation". These results showed the need for a more controlled approach to generating educational content using LLM.

In the second experiment, a systematic feedback validation method was implemented, in which the model response was checked manually or by other LLM. The results showed that even minimal refinement of hints (prompt engineering) can improve the quality of questions by up to 63%. For example, ChatGPT-4 was given an updated instruction indicating the cognitive level: "Formulate a question of the level 'analysis' on the topic of sorting algorithms." In response, the model formulated a question that corresponded to the topic, the complexity level, and the assumed logical reasoning on the part of the student. The resulting question was: "Why is the merge sorting algorithm considered more efficient than the bubble sorting algorithm when working with large amounts of data?"

The third experiment was the most important and introduced an iterative model for improving hints, which allowed LLM to receive structured feedback and, based on it, improve both questions and answer options. The results of the experiment showed that Claude 3.5 Sonnet required the least refinement (a total of 16 iterations for questions and 16 for distractors) and achieved the highest share of valid questions – 92%. ChatGPT-4o showed a stable result: 80% of valid questions with 40 iterations of refinement. Google Gemini showed a significant increase in quality compared to previous experiments (from 53% to 72%), but required the greatest number of clarifications – 108 iterations for questions, 8 for correct answers, and 36 for distractors. In 28% of cases, questions were rejected

after 5 unsuccessful iterations. Typical examples of successful questions were:

Gemini (topic “Data storage”): “Which of the following methods is best suited for storing information about books in the school library?”. Correct answers: database, spreadsheet, CSV file (Comma-Separated Values). Incorrect: a long text file, text documents, or one large file;

ChatGPT-4o (topic “Data transmission”): “How is information transmitted over the Internet?”. Correct: packets, routers, IP addresses. Incorrect: Bluetooth, physical copying, centralised server;

Claude (topic “Wi-Fi”): “Why is the Wi-Fi signal weaker in the school cafeteria than in the computer science classroom?”. Correct answers: physical barriers, number of users. Incorrect: temperature, presence of metal, electricity.

During the study, unsuccessful cases of question generation were identified, highlighting limitations of the models in adhering to both content and formatting requirements for assessment tasks. One such example involved the Gemini PRO 1.5 model attempting to create a multiple-choice question in GIFT format on the topic of operating systems. The model generated the following: `::OS::Select the correct statement about operating systems (=They manage hardware ~They are only used for gaming ~They are optional).` Although the syntax superficially conforms to the GIFT format, the question contains semantic inaccuracies. The option “They are optional” is a misleading distractor since operating systems are fundamental components of most computing devices. Additionally, the instruction implies a single correct answer, while the options may cause confusion regarding the number of correct choices. This example underscores the importance of careful review and editing of AI-generated content before its educational use.

Based on the findings of the experiments, a set of practical recommendations for integrating LLM into the educational assessment process was developed. To ensure alignment with curriculum standards and cognitive depth, it was important to define the target level of complexity in advance, especially using BRT. This helped guide the models toward generating questions that match specific learning outcomes. For instance, when the aim was to create tasks that engage analytical or evaluative thinking, Claude 3.5 Sonnet demonstrated consistent success by producing tasks like evaluating the risks of open Wi-Fi networks or creating examples of phishing emails – both of which reflected higher-order thinking. Another essential aspect was the careful design of structured prompts that included formatting instructions, expected answer types, and restrictions. ChatGPT-4o responded especially well to such instructions, maintaining both formatting accuracy and logical clarity, particularly in tasks that required GIFT or JSON structures. In contrast, the lack of specific constraints often led to format violations and

increased the number of errors, especially in Gemini PRO 1.5 outputs, where correct response separation or hint updates frequently failed. Therefore, it was essential to specify both syntax and structure expectations clearly during the generation process. The iterative clarification and validation process further improved the quality of generated content. Many errors occurred at the initial stage – such as generating incorrect hints or exceeding the complexity boundaries of a task – so manual or semi-automated review loops proved effective. As seen in Table 4, ChatGPT-4o and Claude 3.5 Sonnet significantly improved question validity after such refinement, whereas Gemini PRO 1.5 required more extensive intervention, particularly in clarifying distractors and restructuring JSON elements.

These experiments also highlighted the need for tools that can automatically detect and flag typical issues in AI-generated content. Recurrent problems, such as merging multiple cognitive levels within one task or omitting key syntax elements, could be mitigated by integrated validators in LMS platforms. Automating this process would reduce the need for repeated API calls and teacher interventions, ultimately decreasing time and resource costs. Teachers play a central role in the effective use of these technologies. Providing them with examples of effective prompts and models’ limitations allows for better use of AI in practice. While ChatGPT-4o offered the most balanced combination of speed, accuracy, and formatting stability, Claude 3.5 Sonnet excelled in creating more nuanced tasks at higher cognitive levels. Choosing the model should depend on the intended use case, whether it involves rapid generation of simple test questions or the construction of complex analytical assignments. Ultimately, AI-based systems should not replace educators but serve as intelligent assistants that enhance the efficiency and quality of assessment. When implemented thoughtfully, these tools can support the development of personalized, adaptive testing environments that respond to the real needs of learners, particularly in contexts like Ukrainian distance education, where scalability, clarity, and cognitive relevance are vital.

Analysis of API usage revealed critical insights for practical application, especially in terms of performance stability, cost-efficiency, and technical integration into learning platforms. For instance, ChatGPT-4o demonstrated the fastest response time (average 4.35 seconds) with 81% of correctly generated questions and a moderate API error rate of 5%, making it suitable for real-time classroom use where speed and reliability are essential. In contrast, Google Gemini PRO 1.5 showed a significantly slower response (11.45 seconds on average) and the lowest accuracy (47%), which could limit its effectiveness in environments requiring immediate feedback or large-scale batch generation. Claude 3.5 Sonnet achieved 100% accuracy with a lower API error rate (2%) but required more time per request (8.32 seconds), suggesting it is more

appropriate for generating high-quality assessment items where depth and precision are prioritized over speed. These differences underscore the importance of selecting models based not only on their linguistic capabilities but also on their API performance metrics, especially when integrating AI into platforms like Moodle, where generation time, error frequency, and task complexity directly affect the usability and scalability of the system. Table 6 presents a comparative overview of three LLM (ChatGPT-4o, Claude 3.5 Sonnet, and Gemini PRO 1.5) across four key parameters: accuracy, speed, stability, and refinement capability. This summary helps identify the most suitable model for specific educational needs by clearly illustrating the trade-offs between performance quality and response efficiency.

Table 6. Comparative Evaluation of LLM by Key Parameters

Model	Accuracy (%)	Speed (Avg. response time, s)	Stability (API errors %)	Refinement capability (Valid questions after iteration, %)
ChatGPT-4o	81	4.35	5	80
Claude 3.5 Sonnet	100	8.32	2	92
Gemini PRO 1.5	47	11.45	5	72

Thus, the study identified an effective algorithm for creating adaptive assessment systems using AI. The main components of this algorithm included: generation of questions on computer science topics; quality assessment based on criteria of compliance with the topic, cognitive level, and language; structured feedback; iterative clarification of hints; screening out unsuccessful questions after 5 unsuccessful attempts. This approach allows automating the process of creating tests, reducing the influence of the human factor, and increasing the pedagogical value of digital assessment. Given the rapid spread of distance learning, the proposed algorithm can become the basis for future intelligent educational platforms. Such systems provide self-learning correction, which allows increasing the accuracy and relevance of assessment, and creating more flexible and personalised learning trajectories.

The conducted research revealed significant changes in LMS assessment practices, which were driven by the introduction of sequential AI-based applications. The results showed an increase in the objectivity, efficiency and personalisation of the assessment, which was made possible due to flexible algorithms that adapted to the individual pace of students. The results partially confirmed the observations presented in the study by Kujundziski and Bojadjev [27], which emphasised that AI in education not only transformed learning formats, but also changed the idea of assessment, making it more dynamic and context-dependent. The results partially confirmed the observations presented in the study by Chinta [28], where it was noted that AI in education not

only transformed learning formats, but also changed the idea of assessment, making it more dynamic and context-dependent. The study by Osman and Ahmed [29] analysed the best practices for integrating AI into the learning process, and among them special attention was paid to automated assessment systems that can not only test knowledge, but also support the development of critical thinking skills. This was consistent with the findings regarding the growing role of formative evaluation through AI systems, where feedback became not only instantaneous, but also much deeper in content. Another important trend confirmed by the study results was the ability of AI applications to account for non-fixed assessment parameters, such as metacognitive processes or learning progress, which correlated with the findings of Surve and Londhe [30], who emphasised that AI can analyse and develop non-cognitive skills of students including motivation, perseverance, and self-regulation. This opened up the prospect of creating a more holistic assessment system that went beyond just academic performance. The results of the study were also consistent with those of Pham et al. [31], Cinar et al. [32], who highlighted the positive impact of AI on engineering education, in particular, in the aspect of practical tasks and visualisation of assessment results. The paper emphasized that AI-based systems can better interpret a learner's educational progress than conventional grading scales. Similar dynamics were observed in the results obtained, especially in the context of using graphical reports, cognitive models, and adaptive error analysis. Some aspects of the results found differed from those given by Marmoah et al. [33], where it was noted that the integration of AI with conventional methods is still often formal and does not provide a sufficient level of analytic evaluation results. In a real-world educational environment, AI systems may face imperfect digital infrastructure or teacher resistance, which limits the potential for transformation. As part of the study, a fairly high level of adoption of digital tools among teachers and administrations was recorded, which indicated a gradual overcoming of these barriers.

The implementation of iterative, self-correcting AI applications in this study fundamentally shifts LMS assessment practices from static evaluation to dynamic, personalized adaptation. While prior research has recognized the general utility of AI assistants in improving learning efficiency and student engagement, our findings push this paradigm further by demonstrating that the true pedagogical value of AI lies in its capacity for automated iterative refinement. Unlike standard integrations that simply deploy generative AI for tutoring, our proposed multi-modal pipeline actively validates and self-corrects cognitive tasks against established frameworks like Bloom's Taxonomy [34], [35]. This interpretation suggests that the future efficacy of adaptive learning relies not merely on AI adoption, but on the implementation of structured, AI-driven feedback loops that guarantee the

academic rigor, structural validity, and age-appropriateness of generated content.

The study by Alotaibi [3] found that the integration of AI into LMS in higher education opens up new opportunities for transforming the educational environment, in particular, through adaptive learning and automated assessment. These conclusions were also confirmed by Amer-Yahia [4], who emphasised the role of AI in creating data-driven educational systems to improve the quality of learning. Along with the positive aspects, the study also identified a number of challenges associated with implementing AI in evaluation. For example, research by Arefin [36] focused on ethical issues such as algorithm transparency, bias, and ensuring fairness in evaluation. These aspects were also considered in the paper by Di Marzo Serugendo et al. [37], where an AI assessment scale was proposed for the ethical integration of generative AI into educational assessment. Research has shown that implementing AI in assessment requires new approaches to teacher training. Halkiopoulos and Gkintoni [5] emphasised the need to develop digital competencies among teachers to effectively use AI in teaching and evaluation. These conclusions were also supported in the paper by Komaragiri [38], who reviewed service optimisation strategies using AI to improve data management and evaluation. In the context of Ukrainian higher education, research by Semeniuk et al. [10] demonstrated that the introduction of innovative methodologies and approaches to learning using AI contributed to improving the quality of the educational process. These results were consistent with the findings by Sarnovska et al. [9], who emphasised the effectiveness of using AI in distance learning languages in higher education institutions.

Special attention was paid to the reliability of AI systems when evaluating open-type tasks. The analysis showed that the algorithms mostly showed stable results, but had difficulty recognising the context of culturally labelled statements or unstructured responses. This was partly consistent with the findings by Sriram et al. [39], which pointed to the limited effectiveness of AI in the humanities, where assessment is highly dependent on empathy, interpretation, and social context. However, in technical, medical or mathematical disciplines, AI demonstrated extremely high accuracy and speed, which was confirmed by the results in this study. Interesting parallels can also be drawn with the analysis of the VisioMark system described by Tetteh et al. [40] and Alijoyo [41]. The study showed how automated AI systems can evaluate test tasks with more than 95% accuracy, even in the presence of different fonts, markings, or placeholders. In the course of the study, similar results were observed when using sequential applications for scanning and analysing responses in mixed formats, namely print and digital. Another important aspect of the analysis was the detection of a significant increase in the personalisation of the assessment, which coincided with the results of

Yamijala et al. [42], which emphasised the role of an individual approach based on user behavioural patterns. Effective adaptation of systems to the learning styles of students was recorded, which was manifested in the content, pace, and form of assessment. From the standpoint of interdisciplinary analysis, it should be noted that the results obtained are consistent with the study by Rasool et al. [43], where healthcare innovations were considered. Although the main focus of the study was on oncology, similar principles for analysing large amounts of data and identifying patterns were also applicable in the field of educational assessment. This demonstrated the potential of AI as a tool for a comprehensive analysis of the progress of students. The results of the study were consistent with the trends that dominated modern international research. Sherje [44], for example, proved that the use of AI in the IT field leads to a significant reduction in code verification time, which is similar to the time saved in education through automated evaluation. This emphasised the versatility of automation approaches in various industries [45], [46].

Thus, the results of the analysis showed an increase in the level of personalisation of assessment, adaptation of systems to individual learning styles of students, and the effectiveness of using AI to analyse educational progress. The findings are consistent with interdisciplinary research, particularly in medicine and IT, where similar approaches to big data analysis provided process optimisation. This highlights the prospects of using AI as a tool for automation and personalisation in various industries, including education.

4. Conclusions

An innovative approach was proposed for implementing LLM in education: using a multi-modal structure with consistent LLM challenges to create and validate assessment materials in LMS. Through three progressively improved experiments – ranging from generating algebraic problems in the Moodle GIFT format to creating complex IT/Computer Science questions suitable for high school students that target higher-order cognitive levels – the potential and existing limitations of LLMs in producing reliable, pedagogically sound assessment tasks were demonstrated. Dividing the text generation procedure into stages – first a separate task formulation, and then checking it using external resources – leads to a significant improvement in the correctness of the created questions, and their compliance with the established structure and training requirements. This method provides not only diverse but also high-quality content. Combined with an iterative prompt refinement process and an independent comparative assessment between multiple LLMs, it protects against generation errors, vague responses, and other undesirable artifacts typical of AI text generation. While approaches like Claude 3.5 Sonnet had a near-perfect success rate due

to this improved approach, there are still problems such as high API call latency and cyclical biases in the validation process, as with other generative text evaluation methods, because LLM are used for both generation and validation. For the purposes of validation of automated question generation forms in the educational environment, all created forms were checked in accordance with the BRT. This alignment is particularly important in distance learning, where it is not always possible to provide direct teacher control in real time, and there is a greater need for automated technologies that can effectively and reliably contribute to pedagogical goals.

The study points to a number of methodological limitations, primarily its restricted experimental scope, which focused solely on three specific language models and objective subjects like computer science and algebra, thereby limiting its immediate generalizability to more subjective disciplines. Furthermore, the reliance on LLMs necessitates a rigorous bias analysis, as these models introduce inherent biases from their training data that, when combined with imperfect user inputs, risk generating assessment materials that could alienate diverse student demographics, underscoring the critical need for continuous human-in-the-loop validation to mitigate algorithmic prejudice. Despite these constraints and the increased computational costs from numerous API calls, the study presents sharp practical implications: for educational institutions, this automated pipeline drastically reduces teachers' administrative burden in test creation, allowing them to redirect focus toward personalized mentoring; and for LMS developers, it highlights the urgent need to build intuitive interfaces that abstract prompt-engineering complexities. Ultimately, by proving the viability of a self-improving LLM pipeline, this research lays a solid foundation for scalable, AI-driven tools that enhance both operational efficiency and pedagogical reliability in distance education.

Author Contributions

S.Ye. is the sole author of this manuscript and was responsible for all aspects of the research, including conceptualization, methodology, formal analysis, investigation, writing, and review.

Funding

This research received no external funding.

Data Availability

The author confirms that the data supporting the findings of this study are available in the article.

Acknowledgements

None.

Conflicts of Interest

The author declares no conflict of interest.

References

- [1] A. Alenezi, "Evaluating the effectiveness of AI-powered adaptive learning systems in secondary schools," *International Journal on Studies in Education*, vol. 6, no. 4, pp. 686–700, 2024, doi: [10.46328/ijonse.264](https://doi.org/10.46328/ijonse.264).
- [2] M. Ali, A. Siddique, A. Aftab, M. K. Abid, and M. Fuzail, "AI-Powered customized learning paths: transforming data administration for students on digital platforms," *Journal of Computing & Biomedical Informatics*, vol. 6, no. 2, pp. 195–204, 2024, <https://jcibi.org/index.php/Main/article/view/299>.
- [3] N. S. Alotaibi, "The impact of AI and LMS integration on the future of higher education: Opportunities, challenges, and strategies for transformation," *Sustainability*, vol. 16, no. 23, p. 10357, 2024, doi: [10.3390/su162310357](https://doi.org/10.3390/su162310357).
- [4] S. Amer-Yahia, "Towards AI-powered data-driven education," *Proceedings of the VLDB Endowment*, vol. 15, no. 12, pp. 3798–3806, 2022, doi: [10.14778/3554821.3554900](https://doi.org/10.14778/3554821.3554900).
- [5] C. Halkiopoulos and E. Gkintoni, "Leveraging AI in e-learning: Personalized learning and adaptive assessment through cognitive neuropsychology – A systematic analysis," *Electronics*, vol. 13, no. 18, p. 3762, 2024, doi: [10.3390/electronics13183762](https://doi.org/10.3390/electronics13183762).
- [6] S. Khrapatyi, K. Tokarieva, O. Hlushchenko, O. Paramonova, and I. Lvova, "Research on performance evaluation of higher vocational education informatization based on data envelopment analysis," *STEM Education*, vol. 4, no. 1, pp. 51–70, 2024.
- [7] X. V. Huong, "The implications of artificial intelligence for educational systems: Challenges, opportunities, and transformative potential," *The American Journal of Social Science and Education Innovations*, vol. 6, no. 3, pp. 101–111, 2024, doi: [10.37547/tajssei/Volume06Issue03-17](https://doi.org/10.37547/tajssei/Volume06Issue03-17).
- [8] I.E. Dandachi, "AI-powered personalized learning: Toward sustainable education," in *Navigating the Intersection of Business, Sustainability and Technology*, H. El-Chaarani, I. El Dandachi, S. El Namar, and Z. EL Abiad, Eds., Singapore: Springer, 2024, pp. 109–118, doi: [10.1007/978-981-99-8572-2_5](https://doi.org/10.1007/978-981-99-8572-2_5).
- [9] N. Sarnovska, J. Rybinska, and Y. Mykhailichenko, "Enhancing university remote language learning through innovative applications of artificial intelligence technologies amidst global challenges," *Teaching Languages at Higher Educational Establishments at the Present Stage. Intersubject Relations*, vol. 44, pp. 151–165, 2024, doi: [10.26565/2073-4379-2024-44-10](https://doi.org/10.26565/2073-4379-2024-44-10).
- [10] R. Semeniuk, Y. Holovnia, O. Huda, and N. Savastru, "Innovative methodologies and approaches to teaching with artificial intelligence in Ukrainian Higher Education," *Futurity Education*, vol. 4, no. 1, pp. 24–52, 2024, doi: [10.57125/FED.2024.03.25.02](https://doi.org/10.57125/FED.2024.03.25.02).
- [11] T. K. Vashishth, V. Sharma, K. K. Sharma, B. Kumar, R. Panwar, and S. Chaudhary, "AI-driven learning analytics for personalized feedback and assessment in higher education," in *Using Traditional Design Methods to Enhance AI-Driven Decision Making*, T. V. Nguyen and N. Vo, Eds., London: IGI Global, 2024, pp. 206–230, doi: [10.4018/979-8-3693-0639-0.ch009](https://doi.org/10.4018/979-8-3693-0639-0.ch009).
- [12] A. Jaiswal and C. J. Arun, "Potential of Artificial Intelligence for transformation of the education system in India," *International Journal of Education and Development using Information and Communication Technology*, vol. 17, no. 1, pp. 142–158, 2021.
- [13] O. Karaman, V. Moroz, Y. Bezghynskyi, V. Volodavchyk, and O. Krutko, "Development of the Digital Environment in Ensuring the Quality of Training of Specialists," *Statistics, Optimization and Information Computing*, vol. 15, no. 4, pp. 2727–2742, 2026, doi: [10.19139/soic-2310-5070-3051](https://doi.org/10.19139/soic-2310-5070-3051).
- [14] S. Uzunova, R. Stoyanova, M. Atanassov, and K. Kilova, "Attitudes of Healthcare Service Users in Bulgaria Towards the Application of Teleophthalmology in the Case of Glaucoma," *Healthcare*, vol. 14, no. 2, p. 273, 2026, doi: [10.3390/healthcare14020273](https://doi.org/10.3390/healthcare14020273).

- [15] O. Joseph, E. Dahan, I. Aviv, I. Hadar, E. Bordo, and D. Pezo, "Requirements Engineering for Integrating Quantum Key Distribution with Blockchain Systems," in *2025 IEEE 33rd International Requirements Engineering Conference Workshops (REW)*, Valencia, Spain, pp. 375–382, 2025, doi: [10.1109/REW66121.2025.00055](https://doi.org/10.1109/REW66121.2025.00055).
- [16] A. Artyukhov, N. Artyukhova, A. Navolokina, V. Brych, V. Kozniuk, and L. Kolinets, "Immersive university: Step beyond the metaverse and education," in *Proceedings of the International Conference on Advanced Computer Information Technologies (ACIT)*, pp. 1007–1012, 2025, doi: [10.1109/ACIT65614.2025.11185660](https://doi.org/10.1109/ACIT65614.2025.11185660).
- [17] S. Kiurchev, M. A. Abdullo, T. Vlasenko, S. Prasol, and V. Verkholtantseva, "Automated Control of the Gear Profile for the Gerotor Hydraulic Machine," in *Lecture Notes in Mechanical Engineering*, Cham: Springer, pp. 32–43, 2023, doi: [10.1007/978-3-031-16651-8_4](https://doi.org/10.1007/978-3-031-16651-8_4).
- [18] V. Kyurchev, S. Kiurchev, K. Rezvaya, A. Pastushenko, and S. Glowacki, "Experimental Evaluation of the Impact of the Diametral Clearance on Output Characteristics of a Planetary Hydraulic Motor," in *Lecture Notes in Mechanical Engineering*, Cham: Springer, pp. 84–94, 2023, doi: [10.1007/978-3-031-32774-2_9](https://doi.org/10.1007/978-3-031-32774-2_9).
- [19] R. Shults, A. Urazaliev, A. Annenkov, O. Nesterenko, O. Kucherenko, and K. Kim, "Different approaches to coordinate transformation parameters determination of nonhomogeneous coordinate systems," *Environmental Engineering*, article no. enviro.2020.687, 2020, doi: [10.3846/enviro.2020.687](https://doi.org/10.3846/enviro.2020.687).
- [20] B. Orazbayev, A. Tanirbergenova, K. Orazbayeva, M. Berikbaeva, S. Kaliyeva, L. Kurmagazyeva, and V. Makhatova, "Decision Making for Control of the Gasoline Fraction Hydrotreating Process in a Fuzzy Environment," *Processes*, vol. 12, no. 4, p. 669, 2024, doi: [10.3390/pr12040669](https://doi.org/10.3390/pr12040669).
- [21] E. Dahan, "Shaping Decentralized Collaboration in DAOs with a Requirements Engineering Framework," in *Proceedings of the 2025 IEEE 33rd International Requirements Engineering Conference Workshops (REW 2025)*, Valencia, Spain, pp. 383–394, 2025, doi: [10.1109/REW66121.2025.00056](https://doi.org/10.1109/REW66121.2025.00056).
- [22] A. Amourah, T. Al-Hawary, F. Yousef, and J. Salah, "Collection of Bi-Univalent Functions Using Bell Distribution Associated With Jacobi Polynomials," *International Journal of Neutrosophic Science*, vol. 25, no. 1, pp. 228–238, 2025, doi: [10.54216/IJNS.250121](https://doi.org/10.54216/IJNS.250121).
- [23] E. Ginters and J. C. Revathy, "Hidden and latent factors' influence on digital technology sustainability development," *Mathematics*, vol. 9, no. 21, p. 2801, 2021, doi: [10.3390/math9212801](https://doi.org/10.3390/math9212801).
- [24] E. Ginters, J. M. Gutiérrez, and E. G. Mendiivil, "Mapping of Conceptual Framework for Augmented Reality Application in Logistics," in *2020 61st International Scientific Conference on Information Technology and Management Science of Riga Technical University (ITMS 2020) Proceedings*, pp. 1–6, 2020, doi: [10.1109/ITMS51158.2020.9259302](https://doi.org/10.1109/ITMS51158.2020.9259302).
- [25] I. M. Kadenko, N. V. Sakhno, B. Biró, A. Fenyvesi, R. V. Iermolenko, and O. P. Gogota, "A Bound Dineutron: Indirect and Possible Direct Observations," *Acta Physica Polonica B, Proceedings Supplement*, vol. 17, no. 1, pp. 1A31–1A39, 2024, doi: [10.5506/APhysPolBSupp.17.1-A3](https://doi.org/10.5506/APhysPolBSupp.17.1-A3).
- [26] R. Yermolenko, A. Falko, O. Gogota, Y. Onishchuk, and V. Aushev, "Application of Machine Learning Methods in Neutrino Experiments," *Journal of Physical Studies*, vol. 28, no. 3, 2024, doi: [10.30970/jps.28.3001](https://doi.org/10.30970/jps.28.3001).
- [27] A. P. Kujundziski and J. Bojadjev, "Artificial intelligence in education: Transforming learning landscapes," in *Reimagining Intelligent Computer-Assisted Language Education*, M. Stevkovska, M. Klemenich, and N. Ulutas, Eds., London: IGI Global, 2025, pp. 1–54, doi: [10.4018/979-8-3693-4310-4.ch001](https://doi.org/10.4018/979-8-3693-4310-4.ch001).
- [28] S. Chinta, "The impact of AI-powered automation on agile project management: Transforming traditional practices," *International Research Journal of Engineering and Technology*, vol. 8, no. 10, pp. 2025–2036, 2021, doi: [10.2139/ssrn.5034076](https://doi.org/10.2139/ssrn.5034076).
- [29] S. A. Osman and Z. E. Ahmed, "Navigating AI integration: Case studies and best practices in educational transformation," in *AI-Enhanced Teaching Methods*, Z. E. Ahmed, A. A. Hassan, and R. A. Saeed, Eds., London: IGI Global, 2024, pp. 240–267, doi: [10.4018/979-8-3693-2728-9.ch011](https://doi.org/10.4018/979-8-3693-2728-9.ch011).
- [30] B. C. Surve and B. R. Londhe, "Artificial Intelligence based assessment and development of student's Non-cognitive skills in Professional Education through an online Learning Management System," in *Proc. 4th Int. Conf. on Inventive Systems and Control*, Coimbatore: IEEE, 2020, pp. 329–336, doi: [10.1109/ICISC47916.2020.9171137](https://doi.org/10.1109/ICISC47916.2020.9171137).
- [31] T. Pham, T. B. Nguyen, S. Ha, and N. T. Ngoc, "Digital transformation in engineering education: Exploring the potential of AI-assisted learning," *Australasian Journal of Educational Technology*, vol. 39, no. 5, pp. 1–19, 2023, doi: [10.14742/ajet.8825](https://doi.org/10.14742/ajet.8825).
- [32] O. E. Cinar, K. Rafferty, D. Cutting, and H. Wang, "AI-powered VR for enhanced learning compared to traditional methods," *Electronics*, vol. 13, no. 23, p. 4787, 2024, doi: [10.3390/electronics13234787](https://doi.org/10.3390/electronics13234787).
- [33] S. Marmoah, T. Murwaningsih, F. Nurhasanah, K. Saddhono, A. D. Sutomo, and B. Legowo, "An integration of AI and traditional methodology in the education field in order to: Transform the trends," in *Proc. 4th Int. Conf. on Advance Computing and Innovative Technologies in Engineering*, Noida: IEEE, 2024, pp. 884–889, doi: [10.1109/ICACITE60783.2024.10616679](https://doi.org/10.1109/ICACITE60783.2024.10616679).
- [34] M. Alier, J. Pereira, F. J. García-Peñalvo, M. J. Casañ, and J. Cabré, "LAMB: An open-source software framework to create artificial intelligence assistants deployed and integrated into learning management systems," *Computer Standards & Interfaces*, vol. 92, p. 103940, 2025, doi: [10.1016/j.csi.2024.103940](https://doi.org/10.1016/j.csi.2024.103940).
- [35] S. Banjade, H. Patel, and S. Pokhrel, "Empowering education by developing and evaluating generative AI-powered tutoring system for enhanced student learning," *Journal of Artificial Intelligence and Capsule Networks*, vol. 6, no. 3, pp. 278–298, 2024, doi: [10.36548/jaicn.2024.3.003](https://doi.org/10.36548/jaicn.2024.3.003).
- [36] S. Arefin, "Chronic disease management through an AI-powered application," *Journal of Service Science and Management*, vol. 17, no. 4, pp. 305–320, 2024, doi: [10.4236/jssm.2024.174015](https://doi.org/10.4236/jssm.2024.174015).
- [37] G. Di Marzo Serugendo, M. A. Cappelli, G. Falquet, C. Métral, A. Wade, S. Ghadfi, A.-F. Cutting-Decelle, A. Caselli, and G. Cutting, "Streamlining tax and administrative document management with AI-powered intelligent document management system," *Information*, vol. 15, no. 8, p. 461, 2024, doi: [10.3390/info15080461](https://doi.org/10.3390/info15080461).
- [38] V. B. Komaragiri, "Generative AI-powered service operating systems: A comprehensive study of neural network applications for intelligent data management and service optimization," *Journal of Computational Analysis & Applications*, vol. 33, no. 8, pp. 1841–1856, 2024.
- [39] A. Sriram, K. Ramachandran, and S. Krishnamoorthy, "Artificial intelligence in medical education: Transforming learning and practice," *Cureus*, vol. 17, no. 3, p. e80852, 2025, doi: [10.7759/cureus.80852](https://doi.org/10.7759/cureus.80852).
- [40] D. J. Tetteh, B. P. Okai, and A. N. Beatrice, "VisioMark: An AI-powered multiple-choice sheet grading system," 2023. [Online]. Available: doi: [10.13140/RG.2.2.12796.31368](https://doi.org/10.13140/RG.2.2.12796.31368)
- [41] F. A. Alijojo, "AI-powered deep learning for sustainable industry 4.0 and internet of things: Enhancing energy management in smart buildings," *Alexandria Engineering Journal*, vol. 104, pp. 409–422, 2024, doi: [10.1016/j.aej.2024.07.110](https://doi.org/10.1016/j.aej.2024.07.110).
- [42] S. M. Yamijala, R. M. Chodisetty, C. Chakravorty, and K. P. Sai, "AI-powered learning revolutionizing smart education with personalized learning styles," in *Internet of Behavior-Based Computational Intelligence for Smart Education Systems*, M. Ouaisa, H. Lamaazi, M. El Hamlaoui, and K. Reddy, Eds., London: IGI Global, 2025, pp. 191–212, doi: [10.4018/979-8-3693-8151-9.ch007](https://doi.org/10.4018/979-8-3693-8151-9.ch007).
- [43] S. Rasool, M. Ali, H. M. Shahroz, H. K. Hussain, and A. Y. Gill, "Innovations in AI-powered healthcare: Transforming cancer treatment with innovative methods," *BULLET: Jurnal Multidisiplin Ilmu*, vol. 3, no. 1, pp. 118–128, 2024.
- [44] N. Sherje, "Enhancing software development efficiency through AI-powered code generation," *Research Journal of Computer*

- Systems and Engineering*, vol. 5, no. 1, 2024, doi: [10.52710/rjese.90](https://doi.org/10.52710/rjese.90).
- [45] B. Orazbayev, A. Zhumadillayeva, M. Kabibullin, M. J. C. Crabbe, K. Orazbayeva, and X. Yue, "A Systematic Approach to the Model Development of Reactors and Reforming Furnaces With Fuzziness and Optimization of Operating Modes," *IEEE Access*, vol. 11, pp. 74980–74996, 2023, doi: [10.1109/ACCESS.2023.3294701](https://doi.org/10.1109/ACCESS.2023.3294701).
- [46] L. Isak, O. Babak, and Ye. Hren, "Digital tools in professional education training," *Professional Education: Methodology, Theory and Technologies*, vol. 9, no. 2, pp. 104-125, 2023. doi: [10.31470/2415-3729-2023-18-104-125](https://doi.org/10.31470/2415-3729-2023-18-104-125)