UDC: 37.091:004.8:51

Vasyl Zalizko
Candidate of Physical and Mathematical Sciences,
Doctor of Economic Sciences, Professor, Research Fellow, Artificial Intelligence Centre, Swiss Federal Institute of Technology Zurich (ETH), Zurich, Switzerland
ORCID ID 0000-0001-5362-8270
zwd@ukr.net

AUTOMATED ASSESSMENT OF MATHEMATICAL PROBLEM-SOLVING IN SECONDARY SCHOOL BASED ON THE INTEGRATION OF CHATGPT AND MRIIA+

Abstract. This article presents an approach to automated assessment of school mathematics assessment based on the integration of ChatGPT with the Ukrainian digital educational application Mriia. The proposed approach addresses the growing need for transparent, reproducible, and scalable pedagogical diagnostic tools capable of evaluating not only formal learning outcomes, but also the procedural logic of mathematical reasoning, solution structures, and the development of students' mathematical competencies. The approach is implemented and empirically evaluated using the MRIIA+ prototype, a digital educational platform that enables the automated collection, analysis, and interpretation of educational data within the Mriia application. The institutionalisation of this application has created new opportunities for integrating artificial intelligence tools into formal educational processes. Within the MRIIA+ environment, the integration of ChatGPT and Mriia components supports a systematic three-way information exchange between students, teachers, and parents, as well as the intellectual analysis of educational content. Empirical analysis indicates that the effects of AI-assisted assessment are selective. For the error correction indicator, the automated approach demonstrated statistically significant advantages (Mann-Whitney U, $p_BH < 0.001$, $r \approx 0.4$), whereas for feedback readability and learning engagement, traditional assessment methods remained more effective (Cohen's $d \approx 2.0$). A key feature of the proposed approach is its ability to automatically detect recurring student errors and analyse the structure of mathematical solutions, including handwritten work, while providing personalised feedback in a way that reduces teachers' time expenditure without compromising pedagogical quality. The MRIIA+ prototype proved to be effective in online and blended learning contexts. At the same time, the results indicate that automated assessment performs optimally within a defined homeostatic plateau, representing a balanced interaction between AI-based analysis and the teacher's pedagogical interpretation. This balance enhances assessment objectivity, supports regular learning activities, increases student motivation, and assists teachers in evidence-based decision-making in secondary school mathematics education.

Keywords: ChatGPT; MRIIA+; AI assessment; machine learning; school mathematics; OCR.

1. INTRODUCTION

Current trends in the digitisation of general secondary education in Ukraine mean that we must reconsider how we organise, monitor and assess learning processes, especially when integrating national educational information systems with centralised digital infrastructures such as the Automated Information Complex for Educational Management [1]. This shift is especially significant in mathematics education, where traditional assessment methods often focus on final answers, overlooking students' reasoning processes, problem-solving strategies and the development of mathematical competencies. Moreover, wartime conditions have exacerbated the rapid growth of online and blended learning, increasing the demand for reliable, scalable, and pedagogically sound diagnostic tools that can support continuous monitoring and provide meaningful feedback in digital learning environments.

Modern artificial intelligence technologies open up opportunities to move from fragmentary assessment to systematic analysis of students' learning activities. The use of large

© Vasyl Zalizko, 2026.

language models, such as ChatGPT, allows for the automation of written mathematical solution analysis, the identification of typical errors, the evaluation of the logical structure of arguments, and the formation of detailed feedback. At the same time, the effectiveness of such solutions increases significantly when they are integrated into a formalised digital educational environment that provides for the collection, storage, and interpretation of educational data. In Ukraine, such an environment is provided by the educational application "Mriia" which recently acquired official status and is being implemented within the logic of the state digital ecosystem "Diia". In particular, [1] defines the "Mriia" application as a free, open digital ecosystem designed to support and automate essential elements of the educational process. Its infrastructure enables real-time monitoring of students' learning activities and performance, providing a more objective, data-driven basis for assessment. In this context, integrating ChatGPT with "Mriia" is particularly significant as it enables the exchange of structured data within a unified digital learning environment, supporting the development of advanced diagnostic and feedback mechanisms.

The problem statement. In response to the identified educational needs and methodological challenges, this study advances three core scientific hypotheses addressing the integration of large language models into formal digital educational environments.

Hypothesis H1: To effectively integrate ChatGPT and the Mriya application, it is advisable to create an open-architecture digital learning environment (MRIIA+ [2]) to support regular learning data collection and automated assessment. Such an environment is expected to positively impact learning quality, particularly in mathematics, as reflected in task performance, frequency of error correction, and development of mathematical competence, compared to traditional manual, episodic, result-oriented assessment practices.

Hypothesis H2: Integrating a handwriting recognition module (OCR/handwriting-to-structure) with ChatGPT’s analytical capabilities improves the diagnostic validity of written mathematics assessment by enabling the analysis of not only final answers but also the solution process (sequence of steps, typical errors, and argumentation logic). This integration leads to more accurate identification of learning gaps and a reduction in the proportion of unnoticed or misclassified errors compared to assessment approaches focused solely on results or selective manual review.

Hypothesis H3: Implementing a systematic feedback loop between teachers, students and parents within MRIIA+, based on automated error analysis and personalised recommendations, increases students' motivation to learn and the regularity of their learning activities (e.g. engagement, completion of homework and repeated attempts after making errors) compared to situations in which feedback is delayed, fragmented or not standardised.

Analysis of recent studies and publications. Modern mathematics education research has increasingly shifted from episodic knowledge testing toward data-driven, adaptive, and formative assessment models that emphasise not only final outcomes but also procedural reasoning and feedback quality. Systematic reviews [3], [4] demonstrate that large language models, particularly ChatGPT, can support mathematics education through step-by-step explanations, solution scaffolding, and dialogue-based support, while also opening new opportunities for automated assessment of open-ended responses [5], [6], [7].

At the same time, empirical studies and systematic reviews highlight significant methodological and ethical risks associated with AI-based assessment, including unstable reasoning, plausible but incorrect solutions, and threats to academic integrity [8], [9]. These concerns reinforce the necessity of pedagogically controlled and transparent assessment architectures, especially in school mathematics, where conceptual understanding and procedural accuracy are critical [10]. An increasing amount of research is focusing explicitly on the automated assessment of constructed mathematical responses. Studies based on large-scale datasets have shown that, under controlled conditions, LLM-based scoring can achieve a

level of consistency similar to that of humans, thereby reducing teachers' assessment workload and enabling the provision of scalable formative feedback [11], [12]. However, the quality of automated assessment is strongly influenced by interaction design and prompting strategies, which directly affect the didactic usefulness of AI-generated feedback and prompts [13].

For school mathematics, analysing handwritten solutions is particularly important, as such artefacts contain rich information about procedural logic, typical errors, and reasoning trajectories. Recent work presented at the Educational Data Mining conferences demonstrates the potential of combining OCR-based image processing with LLM-driven analysis to assess handwritten mathematical responses more accurately than result-only approaches [14], [15]. At the same time, methods based on semantic similarity and vector representations, including Sentence-BERT-based approaches [16], exhibit notable limitations. Empirical evidence shows that distance-based models are highly sensitive to genre, context, and data distribution, leading to significant performance degradation when applied across different populations or task types [17]. In educational assessment, this limits their reliability as universal tools for evaluating students' mathematical reasoning.

These findings substantiate the relevance of hybrid assessment models that integrate AI-based analytical capabilities with structured educational environments and human oversight. In this context, the present study builds on prior research in automated mathematics assessment [8], [11], [13], [18], [19] and extends it by focusing on the mutual integration of large language models and structured learning analytics within the MRIIA+ environment.

The article has **a single primary objective**: to provide a scientific justification for and to report the results of a pedagogical experiment on automated assessment in secondary school mathematics under online and blended learning conditions, based on the integration of ChatGPT with the MRIIA digital educational ecosystem. The study focuses on the empirical evaluation of AI-assisted assessment practices, examining their impact on the quality, reliability, and pedagogical usefulness of mathematics assessment in grades 5–8. Particular attention is given to the analysis of handwritten mathematical work, automated error diagnosis, and the role of structured feedback in supporting students' learning processes.

The following research questions are examined in accordance with the objectives and hypotheses outlined above.

RQ1: How does the synergy between ChatGPT and the MRIIA ecosystem within the integrated MRIIA+ environment affect the validity and reliability of mathematics achievement assessment compared to traditional pedagogical assessment methods?

RQ2: To what extent does the integration of optical character recognition (OCR) technologies with large language models enhance the diagnostic accuracy of assessing students' procedural reasoning in geometric problem solving (e.g. calculating the area of a trapezoid or a triangle), compared to assessment approaches based exclusively on final answers?

RQ3: What types and taxonomies of mathematical errors and cognitive difficulties can be automatically identified by the MRIIA+ system, and how can these diagnostics be transformed into adaptive feedback mechanisms for students in grades 5 through 8?

RQ4: What is the measurable effect of automated, AI-generated feedback on students' engagement, error correction behavior, and learning regularity within a digital mathematics learning environment?

RQ5: What pedagogical risks and technical limitations are associated with the implementation of AI-based assessment systems in school education, and which architectural or procedural safeguards are required to ensure the validity, transparency, and academic integrity of automated assessments?

2. RESEARCH METHODS

2.1. Research Model

The study was conducted within an experimental methodological framework combining elements of pedagogical experimentation and learning analytics. This approach involved a comparative analysis of traditional methods of assessing students' academic achievement in mathematics and an automated assessment method implemented through the integration of the ChatGPT language model and the MRIIA digital educational ecosystem within the experimental MRIIA+ environment. The empirical study was conducted over nine months during the 2023/24 academic year, under online and blended learning conditions, which currently constitute the typical instructional context for Ukrainian secondary education during wartime. The extended observation period enabled the immediate effects of introducing automated assessment to be examined, as well as the stability and persistence of changes in students' academic performance and learning behaviour throughout the academic year.

2.2. Experimental base and research sample

The experimental base for the study was the licensed International Online School IIS (iis.org.ua), where digital educational platforms have been systematically used since 2019. Five classes participated in the study, and during the experimental period (2023–2024), 240 students in grades 5–8 were covered, forming two comparable groups:

- experimental group, in which mathematics lessons were assessed using the integrated MRIIA+ environment (ChatGPT + OCR + educational analytics);
- control group, in which teaching and assessment were carried out according to the traditional model (manual checking of written work by the teacher, delayed feedback, lack of automated analysis).

The groups were comparable in terms of student age, curricula, and number of mathematics teaching hours, which ensured correct interpretation of the identified differences as the effect of the applied assessment approach.

2.3. Characteristics of the MRIIA+ digital educational environment

The state application "Mriya" was used as the main communicator with the personal data protection system. It was considered a formalised digital environment for managing the educational process, providing centralised collection and storage of educational data: task completion results, student activity indicators, performance dynamics, and communication between participants in the educational process. To implement mutual integration with the ChatGPT language model, a prototype of the MRIIA+ learning environment was developed based on the open source Moodle code, which supports machine learning and performs the following functions:

- automated collection and structuring of training data;
- transfer of text and graphic materials (scanned or photographed handwritten works) to the analytical modulet;
- obtaining the results of intellectual analysis and returning them to the educational environment in the form of assessments, comments and personalised recommendations.

Thus, MRIIA+ acts as an integration level that ensures technological, methodological, and ethical consistency in the use of generative artificial intelligence in school assessment [2].

2.4. Analysis of handwritten mathematical works in the environment MRIIA+

The research methodology places special emphasis on the analysis of handwritten mathematical work, which remains the primary means of documenting mathematical reasoning and step-by-step problem solving in school practice. In the study, pupils completed written assignments

on paper, after which they photographed or scanned their solutions and submitted them via the MRIIA+ digital educational environment. Consequently, the ChatGPT language model processed authentic handwritten mathematical texts typical for grades 5–8, rather than abstract or artificially generated responses. The processing of handwritten work within the MRIIA+ environment comprised several sequential stages: digitisation of students' solutions, image pre-processing (including quality normalisation and selection of writing areas), recognition of mathematical symbols and spatial structures using an OCR module, transformation of recognised content into a structured representation of solution steps, and subsequent semantic analysis performed by the ChatGPT model.

This approach enabled a shift from evaluating only final answers toward analysing the procedural logic of solutions, including the correctness of transitions between steps, identification of typical errors, omitted reasoning, and logical gaps. Such analysis is essential for formative assessment in mathematics education. Within the MRIIA+ environment, ChatGPT was employed not as an autonomous evaluator but as an analytical component operating under technological and pedagogical control. Its core functions included analysing the structure and sequence of solutions, classifying errors according to predefined criteria, generating explanatory feedback linked to individual solution stages, and producing personalised recommendations to support further learning. To ensure stability and reproducibility of results, interactions with the model were conducted using standardised prompt templates. Overall, ChatGPT functioned as an intelligent analytical module that complemented, rather than replaced, the teacher's professional judgement, enabling a more systematic and efficient analysis of students' handwritten mathematical work.

2.5. System of indicators and evaluation criteria

For empirical verification of the hypotheses and research questions put forward in the work, a multidimensional system of indicators (Table 1) was formed, taking into account the interdisciplinary nature of the study and combining educational, psychological and analytical components. It reflects both the cognitive aspects of learning (quality of solutions, nature and frequency of errors, clarity of explanations) and behavioural and motivational characteristics (regularity of activity, engagement, response to feedback).

Table 1

Description of indicators for assessing the effectiveness of automated assessment

Indicator	What it measures?	Pedagogical content	Why it is important?
Engagement	The degree of active participation of students in the learning process	Reflects academic motivation, interest in the subject, and willingness to engage with the material.	Without active participation, it is impossible to acquire knowledge and develop skills in a sustainable manner.
Error correction	The ability of students to identify and correct their own mistakes	Characterises the development of metacognitive skills, reflection and learning from mistakes.	It is the central mechanism for deep assimilation of mathematical concepts.
Feedback readability	Clarity and accessibility of explanations and comments	Reflects the effectiveness of pedagogical communication between the system and the student	Unclear feedback cannot be used for learning purposes.
Visualisation quality	Effectiveness of graphics l	Promotes the development of intuitive and spatial thinking	Facilitates understanding of abstract mathematical structures
Feedback individualisation	The degree to which explanations are adapted to the characteristics of a particular student	Reflects the level of personalisation of learning	Increases the relevance of learning and supports intrinsic motivation

The MRIIA+ environment can be assessed in terms of its pedagogical effect by using this approach. It also allows for a comparison of traditional and AI assessment in terms of performance and impact on the structure of learning activities and the educational experience of students. The effectiveness of the proposed approach was assessed using a set of quantitative and qualitative indicators. This study separately analysed the impact of the systematic three-way feedback loop between teachers, students and parents, implemented in MRIIA+, on the stability of students' learning activities. The limitations of the study include the dependence of the quality of the analysis on the quality of the digitisation of handwritten works, the limitations of OCR for complex graphic structures, and the need for pedagogical verification of the results of automated assessment.

2.6. Mathematical and statistical research tools and software environment

Mathematical and statistical processing of experimental data was carried out using descriptive and inferential statistics methods, as well as elements of learning analytics. All calculations were performed using the open-source software R, which enables the results to be reproduced within the context of key machine learning principles, thereby ensuring transparency in statistical procedures.

Let the observation matrix have the form

$$X = \{x_{ij}\}, \qquad i = 1, \dots, n,\ j = 1, \dots, m,$$

where x_{ij} – value of the j-th indicator of the educational activity of the i-th student.

The set of indicators included the following key indicators from Table 1:

- *engagement*;
- *error correction*;
- *feedback readability*;
- *feedback individualisation*;
- *visualisation quality*.

For each of these indicators, the following formula was applied:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\,, s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

as well as median, interquartile range (IQR), minimum and maximum values. For each indicator, null and alternative hypotheses were formulated about the absence or presence of differences between AI-assisted and traditional assessment:

$$H_0: \mu_{\text{AI}} = \mu_{\text{Trad}}, H_1: \mu_{\text{AI}} \neq \mu_{\text{Trad}}.$$

Before applying parametric statistical criteria, the assumption of normality of sample distribution was tested using the Shapiro-Wilk test. If the assumption of normality was satisfied, Student's t-test for independent samples was used to compare the mean values; otherwise, the non-parametric Mann-Whitney test was applied.

Given the large number of parallel statistical tests performed for different indicators, the Benjamini–Hochberg procedure was used to correct the significance level (False Discovery Rate, FDR) in order to control the proportion of false positive results. To interpret the practical significance of the identified differences, in addition to the levels of statistical significance, effect size indicators were calculated. In the case of parametric comparisons, Cohen's d coefficient was used, and for non-parametric comparisons, the rank coefficient r was used, which allowed comparing the intensity of effects between different variables.

To account for the hierarchical structure of the data and repeated measurements within the same participants, linear models with mixed effects were used. The general form of the model was as follows:

$$y_{ij} = \beta_0 + \beta_1 \text{AI}_{ij} + \beta_2 \text{Grade}_{ij} + \beta_3 \text{Stream}_{ij} + u_i + \varepsilon_{ij},$$

where y_{ij} – value of the indicator for the i-th student in the j-th measurement, β_k– fixed effects of relevant predictors, u_i – random intercept for a student, reflecting interindividual differences, and ε_{ij}– random error.

2.6.1. Software environment and reproducibility

Statistical analysis and data visualization were performed in the R environment (version 4.3.x) using open source software and standard scientific packages for data processing, statistical analysis, and graphing. The main analysis tools were the tidyverse packages (for data cleaning, transformation and aggregation), janitor (for variable standardisation and data quality control), rstatix (for parametric and non-parametric statistical tests), effectsize (for calculating effect size indicators), ggplot2 (for constructing scientifically sound graphical visualisations) and gt (for generating formatted tables of results). To ensure the reproducibility of the computational experiment, the dependency management system renv was used, which records the exact versions of all packages used and allows the analytical environment to be restored at any time in the future or on another machine. This ensures that all results – statistical estimates, graphs, and tables – can be reproduced from the same data under identical software conditions. The entire analytical process is implemented in the form of sequential, documented «R-scripts» covering the stages of data import, cleaning, statistical assumption testing, test execution, multiple comparison correction, effect size estimation, and visualization construction. This organisation ensures transparency of the methodology, the possibility of independent verification of results, and reuse of the code in future studies.

Thus, the programmatic implementation of the study complies with the principles of open science and reproducibility of results, which is critically important for interdisciplinary research at the intersection of pedagogy, educational measurement, and artificial intelligence.

2.6.2. Program code

This section presents selected fragments of the R program code used for data preprocessing, statistical analysis, and visualisation of the results. The code was developed to ensure transparency, reproducibility, and methodological rigor in the processing of empirical data obtained within the MRIIA+ experimental environment.

Data import and preprocessing. The following code performs data loading, cleaning, filtering, and transformation into a tidy format suitable for subsequent statistical analysis:

```
library(tidyverse)
library(janitor)
df <- read_csv("Stat_UA23-25.csv") %>%
  clean_names()
df_long <- df %>%
  filter(calendar_year == 2024, grade >= 5, grade < 9) %>%
  mutate(
    group = factor(group, levels = c("traditionally", "ai")),
    metric = factor(metric)
  )
```

This step ensures that only observations relevant to the experimental period and the target student population are included and that categorical variables are correctly encoded for inferential analysis.

Statistical testing and effect size estimation. The next block implements the statistical testing pipeline, including distributional diagnostics, nonparametric group comparisons, multiple-testing correction, and effect size estimation:

```
library(rstatix)
```

```
library(effectsize)
  group_by(metric, group) %>%
  shapiro_test(value)
test_results <- df_long %>%
  group_by(metric) %>%
  wilcox_test(value ~ group) %>%
  adjust_pvalue(method = "BH") %>%
  add_significance("p.adj")
effect_sizes <- df_long %>%
  group_by(metric) %>%
  wilcox_effsize(value ~ group)
final_results <- left_join(test_results, effect_sizes, by = "metric")
```

This analytical workflow enables a robust comparison between traditional and AI-supported assessment conditions while controlling for false discovery rates and quantifying the magnitude of observed effects.

3. THE RESULTS AND DISCUSSION

This section presents the results of the pedagogical experiment aimed at evaluating the effectiveness of AI-assisted automated assessment in secondary school mathematics. The findings are organised according to the predefined learning metrics and corresponding research questions. Here and hereafter, to ensure the reproducibility of the analysis and the correct application of statistical models, the data was presented in a normalised format (tidy data, long format), where each row corresponds to one observation "student–indicator" in a specific academic year. Formally, the set of observations is defined as

$$\mathcal{D} = \{(y_k, g_k, r_k, s_k, i_k, m_k)\}_{k=1}^{N},$$

where y_k *–value*, g_k *–calendar_year*, r_k *– grade*, s_k*–stream*, i_k *– student_id*, m_k*–metric*. The group variable (group) coded the learning condition: traditional assessment (traditionally) or AI-assisted assessment (AI).

The normalised data structure allows:

(i) performing consistent comparisons between groups for each indicator;

(ii) taking into account repeated measurements of the same student in different indicators;

(iii) applying mixed-effects models with random student effects and fixed indicator, class and stream effects.

This section presents the results for the 2024 subsample (grades 5–8), which ensures comparability between the experimental and control conditions. Using R, we will form table 2, which reflects the results of comparing the experimental group (AI-assisted assessment) and the control group (traditional assessment) in terms of the main indicators of students' academic performance during the study period. For each indicator, we provide the statistical criterion used, the test statistic value, the initial p-value, the p_BH value adjusted according to the Benjamini-Hochberg procedure, and the corresponding effect size.

Hereinafter, the following abbreviations are used: metric – a quantitative indicator of students' learning activity or feedback quality; test – the statistical procedure applied for comparing the experimental and control groups; statistic – the computed value of the corresponding test statistic; p – the unadjusted significance level; p_BH – the p-value adjusted for multiple comparisons using the Benjamini–Hochberg procedure; effect_type – the standardized measure of effect magnitude corresponding to the applied statistical test, reported as Cohen's *d* for parametric comparisons and as rank-biserial correlation *r* for non-parametric comparisons. Non-parametric Mann-Whitney tests were applied only in cases where the normality assumption was violated, whereas for the feedback readability indicator – whose distributions satisfied the normality assumption – Student's t-test for independent samples was used.

Table 2

Comparison of academic performance between traditional and AI assessment

metric	test	statistic	p	p_BH	effect_type	effect_fmt
engagement	Mann–Whitney U (Wilcoxon rank-sum)	2848.500000	< .001	< .001	r	0.506
error_correction	Mann–Whitney U (Wilcoxon rank-sum)	953.500000	< .001	< .001	r	0.406
feedback_readability	t-test (independent)	4.878466	< .001	< .001	Cohen's d	1.992
visualisation_quality	Mann–Whitney U (Wilcoxon rank-sum)	2332.500000	0.005	0.006	r	0.259
feedback_individualisation	Mann–Whitney U (Wilcoxon rank-sum)	1575.500000	0.236	0.236	r	0.108

The results obtained indicate statistically significant differences between the groups for the indicators of engagement, error correction, feedback readability, and visualisation quality after multiple checks correction (p_"BH" <0.05). No statistically significant differences were found for the feedback individualisation indicator. The greatest practical effect is observed for the feedback readability indicator (Cohen's d=1.99), which corresponds to a very large effect size and indicates a significant advantage of AI-assisted feedback for this parameter. The above results results allowed us to obtain the following visualisation of the relationship between the research parameters (Fig. 1).

Thus, statistically significant differences between traditional and AI assessment for most key indicators of learning activity confirm hypothesis H1 about the pedagogical feasibility of creating an integrated MRIIA+ environment. In particular, the effects identified for the error correction and visualisation quality indicators show that regular automated analysis of learning data provides a more systematic and reproducible assessment compared to traditional episodic approaches. In addition, Figure 1 clearly illustrates the distributions of the values of the main indicators of student learning activity in traditional and AI-assisted groups for 2024. The use of violin diagrams with built-in boxplots and individual observations allows for the simultaneous assessment of the central tendency, variability, asymmetry of distribution, and degree of overlap between groups. This approach provides a more informative view of the data structure compared to the analysis of mean values alone.

The study showed that the engagement indicator, which reflects the overall level of student engagement in learning, demonstrates a systematic shift in distribution towards higher values in the traditional group. The median and interquartile range in this group are higher, while in the AI group there is greater dispersion and concentration of values in the lower range of the scale. This may indicate that traditional forms of assessment currently better support overall learning engagement, particularly in aspects related to social interaction, teacher expectations, and educational practices familiar to students. However, the opposite pattern is observed for the error correction indicator, which reflects the intensity and effectiveness of error correction. The distribution in the AI group is systematically shifted towards higher values, with a greater concentration of observations in the medium-high and high ranges of the scale. This indicates that AI-assisted assessment creates conditions for more frequent and systematic error correction. A likely reason for this is the possibility of prompt, detailed and repeated feedback, which encourages students to actively work on their own mistakes and supports the development of self-correction skills.

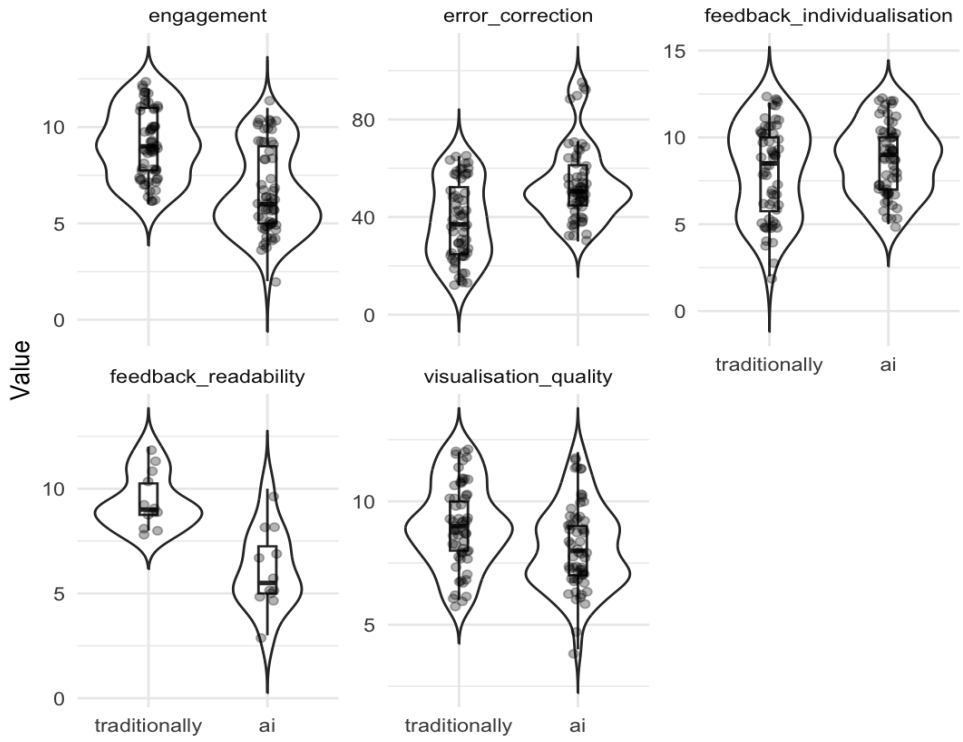

Fig. 1. Comparative analysis of individual observation parameters in the context of using AI to assess students' knowledge of mathematics

For the feedback readability indicator, which characterises the clarity and accessibility of feedback for students, the distributions of the two groups show a marked asymmetry. In the traditional group, the values are significantly higher, and the overlap between the distributions of the two groups is minimal. This indicates that traditional, personalised feedback from teachers is currently perceived by students as more readable and intuitive. At the same time, this result outlines a clear direction for further improvement of AI tools – adapting the linguistic and structural characteristics of automated feedback to students' age and cognitive characteristics.

The visualisation quality indicator, which reflects the quality of the graphic and structural presentation of educational information, is also characterised by higher median values in the traditional group. At the same time, the AI group shows a wider range of values, which may indicate uneven quality of visual solutions or varying degrees of student adaptation to new formats of information presentation. Despite this, the presence of a significant proportion of high values in the AI group indicates the potential of automated systems to scale high-quality visualisations, provided that the design and interfaces are further optimised. Finally, for the feedback individualisation indicator, which reflects the degree of personalisation of feedback, the distributions of the two groups significantly overlap. The medians and interquartile ranges are practically identical, which is consistent with the absence of statistically significant differences between the groups. This result is significant because it shows that, while AI-assisted assessment is currently as personalised as the traditional approach, it does not yet offer a fundamentally higher level of personalised learning feedback.

Thus, Figure 1 clearly demonstrates that the advantages of AI-assisted assessment are selective and are most evident in the procedural aspects of learning, particularly in error correction and support for trial-and-error cycles. At the same time, traditional forms of assessment currently retain their advantage in aspects related to perception, readability, and overall learning engagement. This differentiated picture highlights the usefulness of hybrid

assessment models, which combine the strengths of AI with the pedagogical experience of the teacher.

This demonstrates the advantages of AI-assisted assessment, which are primarily evident in procedural indicators related to error correction. This is consistent with hypothesis H2 and indicates an increase in the diagnostic validity of assessing written mathematical work by analysing not only the final result but also the logic and structure of the solution, which is fundamentally impossible within the framework of traditional testing due to the lack of time and human resource constraints.

A summary of the estimates of the effect sizes of the differences between the traditional and AI-assisted groups is presented in Figure 2, which shows the rank effect size r for non-parametric comparisons and Cohen's d coefficient for parametric comparisons of the feedback readability indicator. The zero vertical line corresponds to no differences between the groups; the distance of the point from zero reflects the practical significance of the effect. The graph allows a quick comparison of which indicators show the most pronounced changes under AI-assisted assessment and which ones have insignificant or no differences.

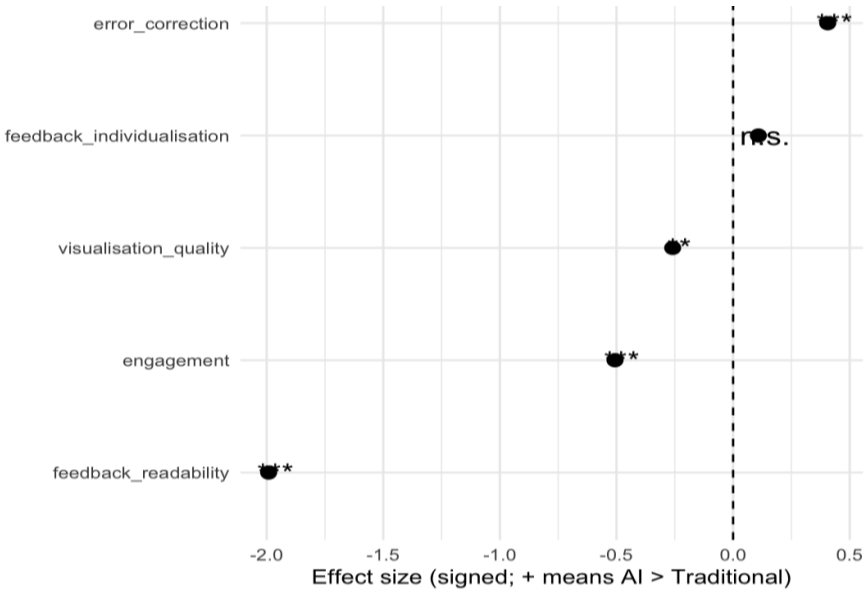

Fig. 2. Overall assessment of the effect of differences between the traditional and AI-assisted groups for each indicator

In Figure 2, the horizontal axis reflects the signed effect size, where positive values correspond to the advantage of AI-assisted assessment, and negative values correspond to the advantage of the traditional approach. The dotted vertical line at zero corresponds to no difference between the groups. The distance of the point from this line is interpreted as a measure of the practical significance of the effect. For the error correction indicator, the most pronounced positive effect of AI-assisted assessment was recorded (rank coefficient r≈0.41, p_"BH" <0.001, ***). This indicates a moderately large effect that has not only statistical but also practical significance. The result suggests that automated assessment tools are more effective in supporting the processes of error detection and correction, probably due to rapid, repeatable, and structured feedback.

For the feedback readability and engagement indicators, there are significant effects in favour of traditional assessment. In particular, a very large effect size was recorded for feedback readability (Cohen's d≈1.99, p_"BH" <0.001, ***), indicating a significant advantage of traditional, personalised feedback in terms of its comprehensibility and perception by students.

This effect size corresponds to a difference of almost two standard deviations between the groups and is unusually large for educational research, emphasising the importance of the human factor in the formation of high-quality explanatory feedback.

For the engagement indicator, an effect in favour of traditional assessment was also found (rank coefficient r≈0.51, p_"BH" <0.001, **). Although the effect size is smaller than for feedback readability, it indicates a stable trend towards higher levels of overall student engagement in traditional settings. This may be due to interpersonal interaction, social expectations, and motivational mechanisms that are more difficult to replicate in automated systems.

For the visualisation quality indicator, a small but statistically significant effect in favour of the traditional approach was recorded (rank coefficient r≈0.26, p_"BH" ≈0.006, *). This effect can be interpreted as small to moderate and indicates that traditional teaching materials currently provide a slightly higher quality visual presentation of information, although the difference is not critical.

Finally, for the feedback individualisation indicator, the effect point is located near the zero line (r≈0.11, p_"BH" =0.236, n.s.), which indicates that there are no statistically significant differences between the groups. This result is significant because it indicates that the level of feedback individualisation in an AI-assisted environment is already comparable to the traditional approach, although it does not yet demonstrate a clear advantage.

In general, Figures 2 and 3 demonstrate the highly differentiated nature of AI-assisted assessment: the most pronounced advantages are observed in the procedural aspects of learning (in particular, in error correction), while in the perceptual and motivational components (readability of feedback, engagement), traditional assessment currently retains its advantage. This pattern of results highlights the value of combining AI tools with teacher support in hybrid assessment models.

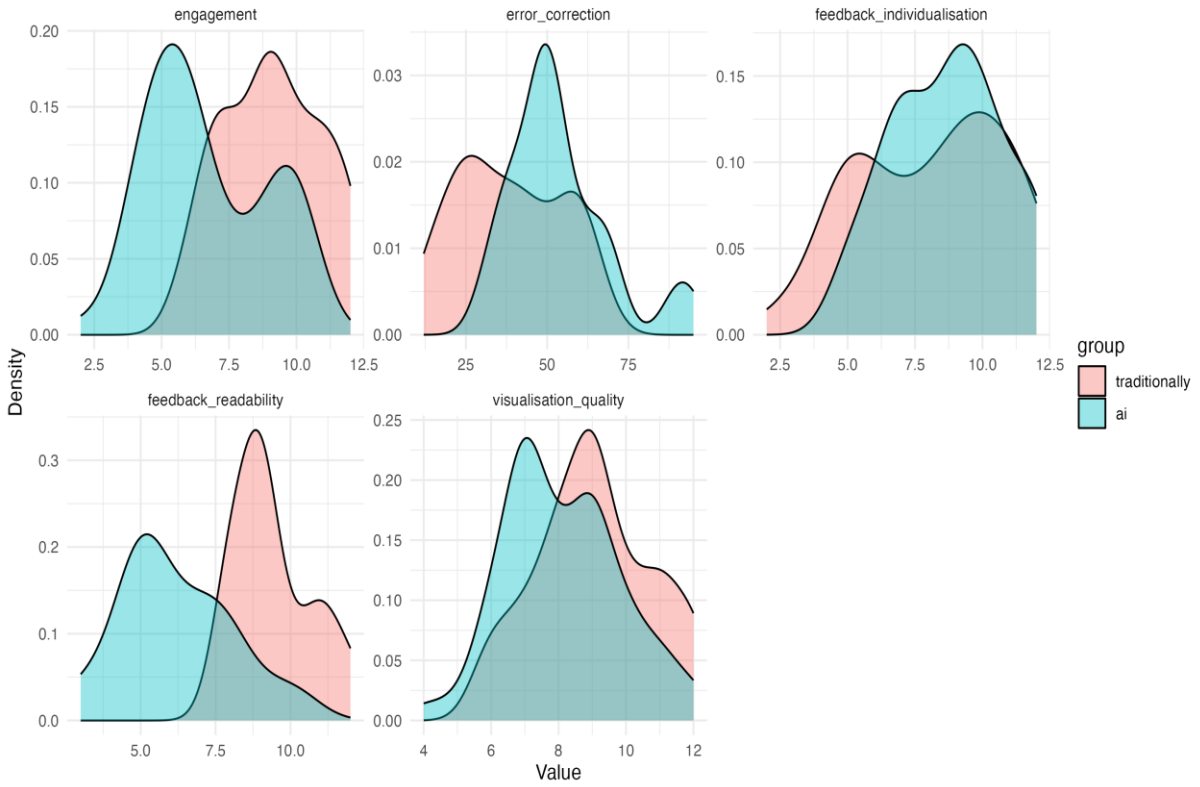

Fig. 3. Interaction between the type of assessment and the type of learning indicator

Thus, we quantitatively differentiated the impact of AI-assisted assessment depending on the type of learning indicator. A moderately large positive effect for error correction confirms the effectiveness of combining OCR and language models for analysing the procedural logic of solving problems (RQ2), while negative effects for feedback readability and engagement indicate the limitations of autonomous AI use without pedagogical interpretation (RQ1).

The results should be interpreted taking into account the organisation of feedback in the experimental group, which studied in online and blended learning environments. In the MRIIA+ environment, feedback is implemented not only in the form of automated comments for the student, but also as a systematic three-way loop "teacher – student – parents", integrated through the educational application Mriia.

Automated analysis of errors and learning activities allows the generation of regular reports on learning progress, which are automatically sent to both students and parents. This reduces the fragmentation of pedagogical control and ensures constant learning feedback beyond the scope of individual lessons or tests.

In this context, the decrease in average engagement scores for AI-assisted assessment should not be interpreted as a clearly negative effect. On the contrary, the results indicate a transition from situational motivation to more regular, structured learning activity, which is consistent with the logic of systemic feedback and confirms the pedagogical feasibility of the tripartite interaction model.

The interaction between the type of assessment and the type of learning indicator has been demonstrated. The intersection of the lines indicates the selective nature of AI's influence: the advantages of automated assessment are most pronounced for indicators related to error correction, while for perceptual and motivational indicators, traditional assessment remains more effective.

Figure 4 depicts a conceptual and analytical model a conceptual and analytical model of a homeostatic plateau – the limits of optimal AI-assisted assessment, which summarises the empirical results presented in the previous sections. The abscissa axis shows the intensity of AI application in the assessment process $A \in [0,1]$, where $A=0$ corresponds to exclusively traditional assessment, and $A=1$ corresponds to full automation of feedback. The ordinate axis shows the normalised educational effect, which allows comparing indicators of different nature on a single scale.

The cross-sectional nature of the interaction between the type of assessment and the type of indicator (Fig. 3 and 4) indicates the selective effectiveness of AI-assisted assessment. This suggests that automated tools are neither better nor worse than traditional ones, but they do have certain advantages in specific areas of educational activity, which directly justifies the need for hybrid assessment models.

From a pedagogical point of view, such results create prerequisites for identifying specific taxonomies of mathematical errors and cognitive difficulties that should be delegated to automated analysis (RQ3), while the interpretative and motivational components of feedback require the active participation of the teacher and family. The identified selectivity of effects also allows us to outline the potential pedagogical risks of excessive automation (RQ5) and emphasises the importance of architectural solutions aimed at verifying and pedagogically interpreting the results of AI analysis.

The processual gain curve reflects the dynamics of procedural learning effects, in particular the effectiveness of error correction. Its shape corresponds to a characteristic saturation growth: in the initial stages of AI implementation, the increase in effectiveness is rapid, but as the share of automation increases, the growth rate gradually decreases, indicating that the limit of usefulness of purely algorithmic intervention has been reached.

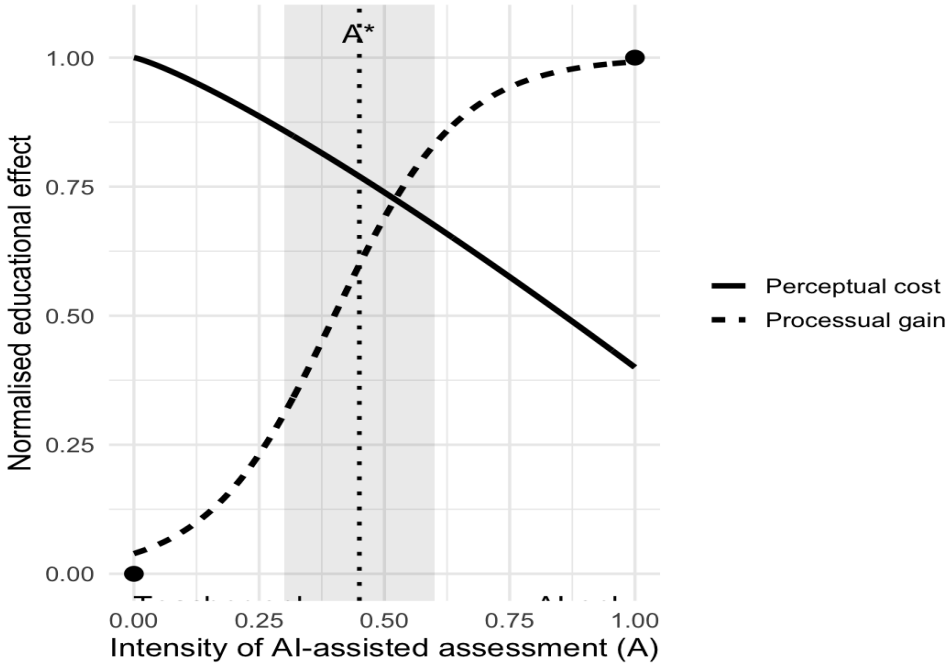

Fig. 4. Homeostatic plateau of the feasibility of using AI when assessing mathematics tasks

This behaviour is consistent with the empirically observed moderately large effect of AI on error correction and reflects the ability of automated systems to perform scalable and repeatable error analysis. Perceptual cost models affect the clarity of feedback and student engagement, which are both important aspects of learning.

Unlike procedural effects, these indicators show a decline with increasing AI intensity, reflecting the gradual loss of elements of interpersonal interaction, contextualisation, and pedagogical interpretation inherent in traditional assessment. This trend is consistent with the significant effects obtained in the study in favour of the traditional approach for feedback readability and engagement.?

The shaded area in the middle of the graph corresponds to a homeostatic plateau – the interval of A values within which the positive procedural effects of AI are already close to saturation, while the negative impact on perceptual and motivational indicators has not yet reached a critical level. It is in this zone that a balance is achieved between the effectiveness of automated analysis and the preservation of pedagogically significant aspects of feedback. The vertical dotted line, which marks the conditional optimum A*, illustrates the concept of a compromise solution rather than a universal maximum of automation.

Thus, Figure 4 demonstrates that the effectiveness of AI-assisted assessment is not a monotonic function of the degree of automation. Instead, it exhibits a distinct optimum that depends on the type of learning indicators considered and on the balance between algorithmic and human components of pedagogical interaction.

The results further indicate that the appropriateness of using AI in educational assessment cannot be reduced to a simple dichotomy of "AI versus teacher." Rather, the empirical evidence points to the selective effectiveness of automated tools, which is expressed primarily in procedural aspects of learning, whereas motivational and perceptual dimensions remain closely linked to the human factor.

In this context, the concept of a homeostatic plateau enables a shift from the comparison of individual indicators toward a more general interpretative framework that explains why and under what conditions AI-based assessment is pedagogically appropriate. This perspective supports the development of hybrid assessment models in which automated systems and teachers fulfil complementary roles rather than replacing one another.

4. CONCLUSIONS AND PROSPECTS FOR FURTHER RESEARCH

The article provides a quantitative analysis substantiating the pedagogical feasibility of using large language models and the MRIIA+ digital educational ecosystem to automatically assess secondary school mathematics lessons. The results demonstrate that the impact of AI assessment is selective and cannot be universally categorised as an increase or decrease in educational quality. Empirical analysis has revealed that procedural indicators of educational activity demonstrate the most pronounced surge in AI effectiveness. Notably, a statistically significant advantage of AI assessment over traditional assessment was found for the error correction indicator (Mann-Whitney $U = 953.5$, $p_BH < 0.001$, $r \approx 0.41$), corresponding to a moderately large effect size. This suggests that the MRIIA+ system has significantly improved its ability to detect and record errors in written mathematical solutions, as well as supporting repeated attempts and students' self-correction.

The opposite trend was observed for the perceptual and motivational components of feedback at the same time. The feedback readability indicator shows that traditional assessment has a significant advantage with a large effect size ($t = 4.88$, $p_BH < 0.001$, Cohen's $d \approx 1.99$), which highlights the important role of human interpretation in creating clear explanations. Similarly, a statistically significant effect in favour of the traditional approach was found for the engagement indicator (Mann-Whitney $U = 2848.5$, $p_BH < 0.001$, $r \approx 0.51$), highlighting the limitations of autonomous AI use in supporting learning motivation. No statistically significant differences between the groups were found for the feedback individualisation indicator (Mann-Whitney $U = 1575.5$, $p_BH=0.236$, $r\approx0.11$), suggesting that AI-assisted feedback achieves a comparable level of individualisation to that of traditional assessment without offering a clear advantage to either approach. This is an important result as it suggests that AI has the potential to be used as a tool for scalable personalisation under pedagogical control.

Generalising the quantitative results enabled us to develop the concept of a homeostatic plateau of AI assessment use. According to this concept, maximum pedagogical effectiveness is not achieved under conditions of full automation, but rather within the range of balanced AI use (Fig. 4). Within this range, an increase in procedural effectiveness (expressed by an effect size of $r\approx0.4–0.5$) is combined with the maintenance of perceptual and motivational indicators at levels similar to those of traditional teaching. An excessive shift towards full automation leads to the saturation of procedural gains and a simultaneous increase in pedagogical losses. This quantitatively confirms the feasibility of hybrid assessment models.

In the context of research questions RQ3–RQ5, the obtained results indicate the possibility of creating taxonomies of common mathematical errors and cognitive challenges through automated analysis. They also highlight the necessity of architectural mechanisms to verify AI assessment results, ensuring academic integrity. The selective nature of the effects of AI allows us to clearly identify areas of pedagogical expediency and risk of automation. In the study conducted, the GPT-4 language model was employed to provide analytical support for automated assessment. It is expected that future generations of large language models (versions 5.x) and agent architectures with long-term memory and contextual adaptation will facilitate the transition from fixed to dynamic modelling of a homeostatic plateau, enabling the optimal level of automation to be adjusted for each student, class and educational topic. The MRIIA+ system's automated detection of recurring mathematical errors and learning difficulties enables the implementation of adaptive feedback, saving time for both students and teachers while automating routine mechanical processes without compromising the quality of instruction and assessment.

Further research should focus on the long-term empirical validation of the proposed hybrid assessment model and on expanding the analytical scope of handwritten mathematical

work. In the present study, the core assessment was conducted in a test-based format within the Moodle environment, while handwritten student solutions were analysed by artificial intelligence only partially and in a simplified manner. This limitation likely reduced the effectiveness of AI-assisted feedback for perceptual and motivational indicators, particularly feedback readability and learning engagement. The integration of ChatGPT with the MRIIA+ ecosystem enhances the validity and diagnostic depth of mathematics assessment by enabling the analysis of both final answers and solution processes through the combined use of large language models and OCR-based handwriting recognition. This approach allows for more precise identification of procedural errors and conceptual misunderstandings compared to traditional result-oriented assessment. At the same time, the implementation of AI-based assessment systems involves pedagogical and technical risks, including over-reliance on automation, algorithmic bias, and potential threats to academic integrity. These risks highlight the necessity of human-in-the-loop supervision, transparent system design, and clear institutional policies to ensure the ethical, pedagogical, and epistemic reliability of automated assessment.

FUNDING

No funding was received for conducting this study.

ACKNOWLEDGEMENT

The author gratefully acknowledges Professor Gerd Kortemeyer (ETH Zürich) for his valuable scientific advice, critical feedback, and methodological recommendations, which substantially contributed to improving the quality of the research, refining the analytical framework, and supporting the interpretation of the results.

REFERENCES (TRANSLATED AND TRANSLITERATED)

[1] Cabinet of Ministers of Ukraine, Issues of functioning of the educational mobile application Mriia, Resolution No. 1334, Oct. 22, 2025. [Online]. Available: https://www.kmu.gov.ua/npas/pytannia-funktsionuvannia-osvitnoho-mobilnoho-dodatka-mriia-k1334. Accessed: Dec. 20, 2025. (in Ukrainian)

[2] V. Zalizko, "MRIIA+: An innovative learning platform for grades 0–12, powered by IIS-GPT," Online educational platform, 2019. [Online]. Available: https://mriia.school-top.com. Accessed: Dec. 20, 2025. (in English)

[3] M. Pepin *et al.*, "A scoping survey of ChatGPT in mathematics education," *Digital Experiences in Mathematics Education*, 2025. https://doi.org/10.1007/s40751-025-00172-1. (in English)

[4] H. S. Almarashdi *et al.*, "Unveiling the potential: A systematic review of ChatGPT in transforming mathematics teaching and learning," *Eurasia Journal of Mathematics, Science and Technology Education*, vol. 19, no. 1, 2024. https://doi.org/10.29333/ejmste/15739. (in English)

[5] Y. Wardat *et al.*, "ChatGPT: A revolutionary tool for teaching and learning mathematics," *Eurasia Journal of Mathematics, Science and Technology Education*, vol. 19, no. 4, 2023. https://doi.org/10.29333/ejmste/13272. (in English)

[6] M. Turmuzi *et al.*, "ChatGPT in school mathematics education: A systematic review of opportunities, challenges, and pedagogical implications," *Teaching and Teacher Education*, vol. 170, Art. 105286, 2026. https://doi.org/10.1016/j.tate.2025.105286. (in English)

[7] L. Yan *et al.*, "Practical and ethical challenges of large language models in education: A systematic scoping review," *British Journal of Educational Technology*, 2024. https://doi.org/10.1111/bjet.13370. (in English)

[8] B. Pepin, *et al.*, "Mathematics education in the era of ChatGPT: Investigating its meaning and use for school and university education – Editorial to special issue," *Digital Experiences in Mathematics Education*, vol. 11, pp. 1–8, 2025. https://doi.org/10.1007/s40751-025-00173-0. (in English)

[9] H. Bastani et al., "Generative AI without guardrails can harm learning: Evidence from high school mathematics," *Proceedings of the National Academy of Sciences of the United States of America*, 2025. https://doi.org/10.1073/pnas.2422633122. (in English)
[10] A. Rasila *et al.*, "On automatic assessment and conceptual understanding," *Teaching Mathematics and its Applications*, vol. 34, no. 3, pp. 149–159, 2015. https://doi.org/10.1093/teamat/hrv013. (in English)
[11] S. Baral *et al.*, "Automated assessment in math education: A comparative analysis of LLMs for open-ended responses," in *Proc. EDM'24*, 2024. https://doi.org/10.5281/zenodo.12729932. (in English)
[12] W. Morris *et al.*, "Automated scoring of constructed response items in math assessment using large language models," *International Journal of Artificial Intelligence in Education*, 2024. https://doi.org/10.1007/s40593-024-00418-w. (in English)
[13] S. Schorcht *et al.*, "Prompt the problem: Investigating the mathematics educational quality of AI-supported problem solving by comparing prompt techniques," *Frontiers in Education*, vol. 9, Art. 1386075, 2024. https://doi.org/10.3389/feduc.2024.1386075. (in English)
[14] S. Baral *et al.*, "Auto-scoring student responses with images in mathematics," in *Proc. EDM'23*, 2023. https://doi.org/10.5281/zenodo.8115645. (in English)
[15] S. Baral *et al.*, "Automated feedback in math education: A comparative analysis of LLMs for open-ended responses," arXiv:2411.08910, 2024. https://doi.org/10.48550/arXiv.2411.08910. (in English)
[16] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. EMNLP*, 2019. https://doi.org/10.48550/arXiv.1908.10084. (in English)
[17] M. Kocher and J. Savoy, "Distance measures in author profiling," *Information Processing & Management*, vol. 53, no. 5, pp. 1103–1119, 2017. https://doi.org/10.1016/j.ipm.2017.04.004. (in English)
[18] S. Poka and C. Herman, "Automatic assessment of mathematics," in *Proc. International Conference on Information Technology and Learning Tools*, 2017, pp. 111–118. https://doi.org/10.12753/2066-026X-17-190. (in English)
[19] G. Kortemeyer *et al.*, "Artificial-intelligence grading assistance for handwritten calculus exam," arXiv:2510.05162, 2025. https://doi.org/10.48550/arXiv.2510.05162. (in English)

Text of the article was accepted by Editorial Team *05.01.2026p.*

АВТОМАТИЗОВАНЕ ОЦІНЮВАННЯ РОЗВ'ЯЗУВАННЯ МАТЕМАТИЧНИХ ЗАДАЧ У СЕРЕДНІЙ ШКОЛІ НА ОСНОВІ ІНТЕГРАЦІЇ CHATGPT ТА MRIIA+

Василь Залізко
кандидат фізико-математичних наук, доктор економічних наук,
професор, науковий співробітник,
Центр штучного інтелекту Федеральної вищої технічної школи
Цюриха (ETH), м. Цюрих, Швейцарія
ORCID ID 0000-0001-5362-8270
zwd@ukr.net

Анотація. У статті запропоновано підхід до автоматизованого оцінювання уроків математики на основі інтеграції великої мовної моделі GPT та українського освітнього застосунку «Мрія». Актуальність дослідження зумовлена зростаючою потребою в прозорих, відтворюваних і масштабованих інструментах педагогічної діагностики, здатних враховувати не лише формальні результати навчання, а й логіку математичних міркувань, структуру розв'язань і динаміку формування навчальних компетентностей учнів. Запропонований підхід реалізується та оцінюється за допомогою прототипу MRIIA+, цифрової освітньої платформи, яка дозволяє автоматично збирати, аналізувати та інтерпретувати освітні дані в межах додатка «Мрія». Інституціоналізація цього додатка відкрила нові можливості для інтеграції інноваційних інструментів штучного інтелекту у формалізовані освітні процеси. Взаємна інтеграція компонентів GPT та «Мрія» у середовищі MRIIA+ забезпечила системний тристоронній обмін інформацією між учнями, вчителями та батьками, а також інтелектуальний аналіз навчального контенту. Емпіричний аналіз показав селективний характер ефектів AI-оцінювання: для досліджуваного показника «виправлення помилок» зафіксовано статистично значущу перевагу автоматизованого підходу (Mann-Whitney U, $p_{BH} < 0.001$, $r \approx 0.4$), тоді як для «читабельності зворотного зв'язку» та «навчальної залученості» перевага зберігається за традиційним оцінюванням (Cohen's $d \approx 2.0$). Ключовою особливістю запропонованого підходу є його здатність автоматично виявляти повторювані помилки учнів і аналізувати структуру математичних рішень, включно

з рукописною роботою, водночас надаючи персоналізований зворотний зв'язок, зменшуючи витрати часу вчителів без втрат для якості педагогічної роботи. Створений прототип інтегрованої системи MRIIA+ довів свою ефективність в умовах онлайн- та змішаного навчання. При цьому максимальна ефективність автоматизованого оцінювання досягається в межах окресленого гомеостатичного плато – збалансованого рівня взаємодії між AI-аналізом і педагогічною інтерпретацією вчителя, що підвищує об'єктивність оцінювання, регулярність навчальної діяльності, мотивацію учнів та підтримує прийняття обґрунтованих педагогічних рішень у середній школі.

Ключові слова: ChatGPT; Мрія+; AI-оцінювання; машинне навчання; шкільна математика; OCR.

This work is licensed under Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.