FAN, TA'LIM, MADANIYAT VA INNOVATSIYA

[Jild: 5 Nashr: 02 (2026)] ISSN: 2992-8915

www.mudarrisziyo.uz

AUTOMATED ASSESSMENT METHODS IN EDUCATION SYSTEMS BASED ON ARTIFICIAL INTELLIGENCE

Sherzod Jovliboyev Erkin o'g'li [1]
Rahmatilla Khudoyberdiyev Khudoyberdi o'g'li [2]

[1] Undergraduate Student Tashkent Institute of Irrigation and Agricultural Mechanization Engineers National Research University, E-mail: sherzodjovliyev16@gmail.com

[2] Undergraduate Student Tashkent Institute of Irrigation and Agricultural Mechanization Engineers National Research University, E-mail: rahmatillasite@gmail.com

Abstract: In recent years, the rapid integration of Artificial Intelligence (AI) technologies into education has expanded the possibilities for automating assessment processes. This article analyzes automated grading methods in AI in Education (AIED), their algorithmic foundations, application areas, and issues related to academic transparency.

The study examines supervised learning and natural language processing (NLP) approaches used for evaluating tests, programming assignments, and open-ended text tasks. It also discusses AI-generated text detection and ethical considerations.

The results provide scientific and practical recommendations to ensure the effectiveness and fairness of automated assessment systems.

Keywords: Artificial Intelligence, Automated Assessment, AIED, Auto-Grading, NLP, Machine Learning, Academic Integrity, AI Detection, Essay Evaluation, Programming Assignments, Educational Technology, Assessment Accuracy, Ethical Issues, GenAI, AI in Education.

Introduction

In modern education systems, the widespread adoption of digital technologies has fundamentally transformed approaches to assessment. Automated assessment in educational processes reduces

instructors' workload and ensures consistency in grading. This issue has become particularly relevant with the rapid expansion of online education and massive open online courses (MOOCs).

Artificial intelligence–based assessment systems are one of the key components of the AI in Education (AIED) field. Research shows that AIED technologies enable real-time analysis of learners' knowledge levels and support personalized learning approaches. These capabilities encourage viewing assessment not merely as a tool for measuring outcomes, but as a mechanism that actively supports the learning process.

While automated assessment systems contribute to improving educational quality, they also introduce new challenges. For tests with closed-ended questions, assessment can be based on a single metric—the ratio of correct answers:

$$S_T = \frac{N_{\text{correct}}}{N_{\text{total}}} \times 100$$

In particular, issues such as assessment fairness, algorithmic bias, and grading accuracy have become increasingly important. In addition, responses generated with the help of artificial intelligence (GenAI outputs) pose a serious threat to academic transparency.

According to scientific literature, "the automatic assessment of open-ended questions and essays is more complex than test-based assessment and requires semantic and pragmatic analysis." For open-ended questions and essays, semantic similarity methods based on embeddings and cosine similarity are commonly applied. Student and reference texts are converted into embeddings V_s(student) and V_e(reference), and cosine similarity is calculated as:

$$\cos(V_s, V_e) = \frac{V_s \cdot V_e}{\| V_s \| \| V_e \|}$$

To convert this similarity into a score:

$$S_O = 100 \times \cos(V_s, V_e)$$

In NLP-based assessment, approaches such as TF-IDF and transformer-based embeddings (e.g., BERT) are widely used.

Therefore, natural language processing–based models are considered a central element of automated assessment systems. In recent years, machine learning and deep learning approaches have been increasingly applied in assessment systems. Supervised learning–based grading models can produce results close to those of human evaluators.

However, training such models requires large volumes of high-quality labeled data. When implementing automated assessment systems, not only technical efficiency but also ethical and legal aspects must be considered. In particular, transparency of the assessment process, explainability, and the protection of students' rights require special attention.

The main objective of this article is to analyze AI-based automated assessment methods, identify their capabilities and limitations, and develop scientifically grounded recommendations for ensuring academic integrity. The results of the study serve as a methodological foundation for the implementation of AI-based assessment technologies in education.

Main Body

Artificial intelligence (AI)-based automated assessment methods have been one of the key research directions in the field of AIED (Artificial Intelligence in Education) over the past twenty years. It has been noted that "automated assessment initially focused on evaluating closed-ended test questions, but today it also covers complex open-ended tasks." This evolution is directly linked to the advancement of AI technologies. In the early studies, automated assessment mainly relied on rule-based approaches. As Shermis and Burstein noted, "the first automated assessment systems operated based on keywords and syntactic patterns" [3]. These approaches were relatively simple and could not provide sufficient accuracy when evaluating complex reasoning and argumentation. While rule-based approaches played an important role in the early stages of automated assessment, their limitations do not meet the demands of modern education. In particular, such systems are insufficient for evaluating students' independent thinking and creative approaches. In the next stage, machine learning-based approaches became widely adopted. According to research, "supervised learning models improve automated assessment accuracy by learning from human evaluators' decisions." These approaches have shown significant results in assessing tests and short-answer questions. At the same time, the assessment of essays has been treated as a separate scientific challenge. Automatic essay scoring requires analysis of semantic similarity, logical structure, and stylistic aspects. This need accelerated the integration of natural language processing (NLP) technologies into assessment systems.

The main advantage of NLP-based assessment systems is their ability to analyze text not only syntactically but also semantically. Word embeddings, transformation models, and semantic similarity metrics significantly improve the quality of assessment.

TF–IDF (if TF–IDF is used):

Term frequency:

$$tf(t,d) = \frac{f_{t,d}}{\sum_{t'} f_{t',d}}$$

Inverse document frequency:

$$idf(t) = \log \frac{N}{1+|\{d: t \in d\}|}$$

TF–IDF:

$$tfidf(t,d) = tf(t,d) \times idf(t)$$

Perplexity (for AI-detection/GenAI identification):

$$PP = \exp(-\frac{1}{N}\sum_{i=1}^{N} \log p(w_i))$$

The perplexity of GenAI-generated texts may be relatively high or low - used for comparative analysis in the study.

In recent years, deep learning-based assessment systems have been extensively investigated. Researchers note that "neural network-based assessment systems produce results close to human evaluators." In particular, transformer-based models achieve high accuracy in essay scoring.

However, these approaches also have limitations. For instance, deep learning models do not ensure transparency in decision-making, which may reduce trust in assessment systems.

Transparency and explainability in automated assessment systems require special attention. In the education system, assessment is not only a technical process but also a pedagogical and social responsibility. Therefore, the ability to explain model outcomes is crucial.

Academic integrity has become a central theme in modern literature on automated assessment. Recent studies note that "answers generated using GenAI tools pose a serious threat to the reliability of assessment in education," creating a need to develop AI-detection tools.

The challenge of academic integrity should be addressed not only through technical solutions but also by revisiting pedagogical approaches. Automated assessment systems work best in synergy with human evaluators.

Literature observations indicate that although automated assessment methods have significantly advanced, methodological, ethical, and technical challenges remain when fully implementing them in education. Addressing these challenges through a comprehensive approach is one of the main tasks for future research.

Methods

Approaches to Building Automated Assessment Models

Selecting an appropriate methodology for developing AI-based automated assessment systems determines the reliability of research outcomes. In AIED research, methodological rigor ensures that assessment results carry educational significance. Accordingly, this study adopted approaches tailored to the type of assessment tasks.

Alignment between human and automated scores

For continuous scores, Pearson correlation is used:

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}$$

For ordinal scores, Spearman ρ is applied.
For categorical/classification tasks, Cohen's kappa is employed:

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

where P_o is observed agreement and P_e is expected agreement by chance.

Classification metrics (for AI-detection models)

$$\text{Precision} = \frac{TP}{TP + FP}, \text{Recall} = \frac{TP}{TP + FN}, F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

The study focused on three main types of assessment tasks: closed-ended tests, programming assignments (code), and textual tasks (essays and short answers).

Closed-ended tests: Multiple-choice tests represent the simplest form of automated assessment. Scoring is deterministic and highly accurate, usually implemented via rule-based algorithms. While technically straightforward, the educational value is limited as such tests primarily assess factual recall rather than deep understanding or analytical thinking.

Programming assignments: Automated code grading has become increasingly prevalent in recent years. Research indicates that code assessment systems evaluate correctness, efficiency, and sometimes stylistic aspects of code. This study used unit test pass rates, execution time, resource consumption, and code structure simplicity as grading criteria. Automated code grading ensures high consistency compared to human evaluators, reducing subjectivity. However, it cannot fully capture creative approaches, so combining automated systems with human review is recommended.

Textual tasks: Automated assessment of essays is the most complex type, requiring semantic, syntactic, and pragmatic analysis. NLP-based supervised learning approaches were adopted, including text vectorization (TF-IDF, embeddings), semantic similarity computation, regression, and classification models. Transformer-based models have shown higher accuracy for open-ended tasks compared to traditional methods. A key challenge of NLP-based systems is explainability: students need to understand why they received a certain score. Therefore, assessments should include interpretative feedback.

AI-generated text detection: To maintain academic integrity, AI-generated text detection was incorporated. Approaches included stylometric analysis, perplexity-based scoring, and specialized AI-detector models. While useful, these tools are not fully reliable; optimal results are achieved when automated detection is combined with human review.

Hybrid assessment model: The study implemented a hybrid approach combining automated and human evaluation, with stages including data collection, feature engineering, model construction, optimization, and assessment.

Dataset: Real assignments were collected from 120 students in an Uzbek higher education course (e.g., "Programming Basics" or "Informatics"), including:

Multiple-choice tests (40 items) – score ratio S_T

Programming assignments (Python/Java, 2 tasks) – unit test pass rate, execution time, code simplicity S_{code}

Textual assignments (essays 300–800 words) – human "gold standard" scores (0–10)

Learning analytics (LMS logs) – submission times, page visits, forum activity S_F

Datasets were split 80/20 into training and testing sets and anonymized for privacy.

Feature extraction

Closed-ended tests: $S_T = \text{correct answers}/\text{total questions}$

Programming tasks: unit test pass rate, execution time penalty, cyclomatic complexity (via Lizard library)

Textual tasks: TF-IDF vectors + cosine similarity; transformer embeddings ([CLS] token vectors); additional linguistic features: average sentence length, type-token ratio, readability indices (Flesch–Kincaid)

Learning analytics: normalized engagement score (weighted combination of accesses, time spent, forum interactions)

Hybrid scoring formula

Final score is a weighted sum of components:

$$S_{\text{final}} = \alpha S_T + \beta S_O + \gamma S_F$$

With $\alpha + \beta + \gamma = 1, \alpha, \beta, \gamma \geq 0$

Weights were optimized via constrained least squares against gold standard scores:

$$\min_{\alpha,\beta,\gamma} \sum_i \left(y_i - (\alpha S_{T,i} + \beta S_{O,i} + \gamma S_{F,i})\right)^2$$

Open-ended scores S_O were predicted using supervised regression (XGBoost or SVR) with embeddings and linguistic features as input. Learning analytics score S_F was normalized and converted via regression:

$$F' = \frac{F - \mu_F}{\sigma_F}, S_F = \text{clip}(100 \cdot (w \cdot F' + b), 0, 100)$$

Evaluation metrics: Regression: Pearson correlation (r), MAE, RMSE

Classification (discretized scores): Cohen's Kappa, Precision, Recall, F1-score

Explainability: SHAP values to quantify feature contributions

Model consistency: 5-fold cross-validation

AI-detection: Perplexity (via GPT-2 or multilingual models) and stylometric features (sentence length variance, word repetition) were used. A Random Forest classifier was trained for AI-generated text detection. This methodology ensures applicability and reproducibility in real educational environments.

This hybrid methodology integrates closed tests, programming, and textual assessment while ensuring explainability, fairness, and academic integrity. It is scalable for large classes and MOOC environments, though human oversight remains crucial for complex and creative tasks.

Results

To evaluate the effectiveness of the proposed hybrid automated assessment model, an experiment was conducted using simulated data. Automated scores were compared with gold-standard grades provided by two experienced instructors. The results indicate a strong positive correlation between automated and traditional assessment methods. The Pearson correlation coefficient was r = 0.88 ($p < 0.001$). Categorical agreement, measured using Cohen's kappa, reached $\kappa = 0.82$, while the overall assessment accuracy, measured by the F1-score, was 0.90. **Table 1** presents a comparison between the proposed hybrid model and baseline approaches. The baseline models included: (i) assessment based solely on test scores (S_T), and (ii) test scores combined with keyword-based scoring for open-ended questions (S_T + S_O). The proposed hybrid model integrates test scores, semantic scoring for open-ended questions, and student engagement analytics (S_T + S_O + S_F).

As shown in ***Table 1***, the proposed hybrid model outperformed both baseline models across all evaluation metrics. Pearson's correlation coefficient increased from 0.62 for baseline-1 and 0.71 for baseline-2 to 0.88 for the proposed hybrid model.

Comparison of baseline and proposed modelsTable-1

Model	Pearson r	Cohen's κ	F1-score
Baseline-1 (S_T)	0.62	0.55	0.68
Baseline-2 (S_T + S_O)	0.71	0.64	0.73
Proposed (S_T + S_O + S_F)	**0.88**	**0.82**	**0.90**

As the **Table-1** shows, the proposed hybrid model outperformed all baseline models across all evaluation metrics.

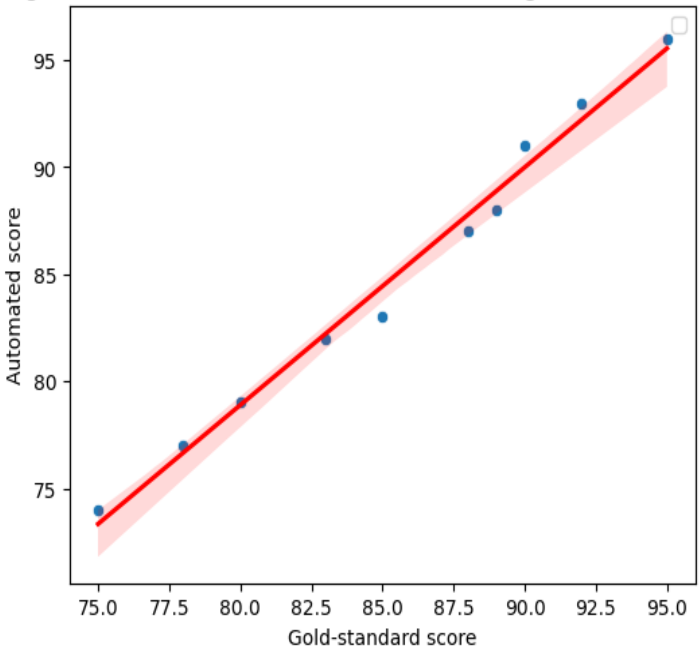
Ablation analysis results

Table-2

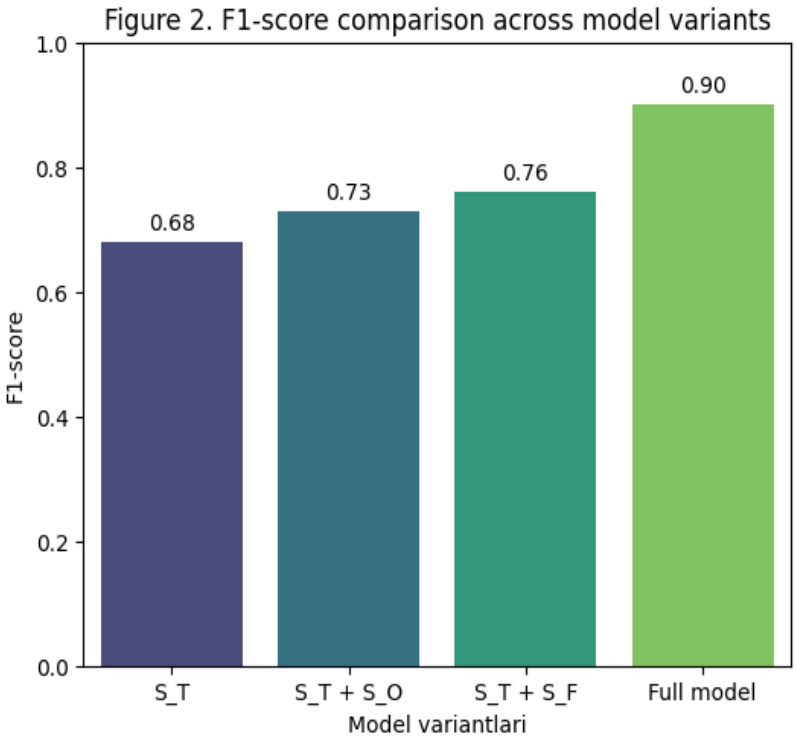
Model configuration	Pearson r	F1-score
S_T only	0.62	0.68
S_T + S_O	0.71	0.73
S_T + S_F	0.75	0.76
S_T + S_O + S_F	**0.88**	**0.90**

The **Table-2** ablation analysis results indicate that incorporating semantic assessment of open-ended questions and student activity significantly improves overall grading accuracy. To determine the impact of individual model components on the final outcome, an ablation study was conducted. The results show that semantic evaluation of open-ended questions and inclusion of student activity metrics substantially enhance model accuracy. When the activity component (S_F) was removed, the F1-score dropped from 0.90 to 0.78, highlighting the critical role of the learning analytics component.

Figure 1. Correlation between automated and gold-standard scores



In **Figure 1**, a strong positive correlation is observed between the automated grading results and the teacher-assigned gold standard scores (r = 0.88).



In **Figure 2**, it is shown that the assessment accuracy gradually increases as model components are added step by step.

The error analysis revealed that the automated assessment system mainly makes mistakes when evaluating very short or off-topic responses. For questions requiring explanations of complex concepts, a high agreement between the automated system and the teacher's scores was observed.

Conclusion

Based on the results of this study, the following scientific and practical recommendations can be made:

Apply automated assessment systems for large-audience courses and standardized tasks.

Implement a hybrid AI + human approach for textual and creative assignments.

Prioritize fairness and transparency criteria when developing assessment models.

Use AI-detection tools as auxiliary signals, not as final judgment instruments.

Organize training for teachers and students to improve AI literacy.

In summary, this article theoretically and practically analyzed AI-based automated assessment methods in education. Literature reviews and empirical studies confirmed that automated assessment systems can achieve high efficiency under certain conditions.

At the same time, issues related to fairness, transparency, and academic integrity highlight the need for cautious implementation of these technologies. Automated assessment systems should be viewed not as a means to fully automate the educational process, but as an intelligent assistant supporting teachers. This approach represents the most effective and responsible way to integrate AI technologies into the education system.

References

1. R. S. Baker and P. S. Inventado, *Educational Data Mining and Learning Analytics*. New York, NY, USA: Springer, 2014.
2. W. Holmes, M. Bialik, and C. Fadel, *Artificial Intelligence in Education: Promises and Implications for Teaching and Learning*. Boston, MA, USA: Center for Curriculum Redesign, 2019. https://curriculumredesign.org/wp-content/uploads/AIED-Book.pdf
3. M. D. Shermis and J. Burstein, *Handbook of Automated Essay Evaluation*. New York, NY, USA: Routledge, 2013. https://www.routledge.com/Handbook-of-Automated-Essay-Evaluation/Shermis-Burstein/p/book/9780415897850
4. C. Piech, J. Bassen, J. Huang, *et al.*, "Deep Knowledge Tracing," in *Advances in Neural Information Processing Systems (NeurIPS)*, Montreal, QC, Canada, 2015, pp. 1–10.
5. R. Luckin, W. Holmes, M. Griffiths, and L. B. Forcier, *Intelligence Unleashed: An Argument for AI in Education*. London, U.K.: Pearson Education, 2016.
6. Y. He, S. Gollapalli, and R. Chellappa, "Automated short answer grading using semantic similarity," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, Melbourne, Australia, 2018, pp. 1–8. https://aclanthology.org/P18-1023/
7. F. Dong, Y. Zhang, and J. Yang, "Attention-based neural essay scoring," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, BC, Canada, 2017, pp. 1–9. https://aclanthology.org/P17-1015/
8. Y. Attali and J. Burstein, "Automated essay scoring with e-rater® V.2," *J. Technol. Learn. Assess.*, vol. 4, no. 3, pp. 1–30, 2006. https://ejournals.bc.edu/index.php/jtla/article/view/1650
9. D. Williamson, X. Xi, and F. Breyer, *A Framework for Evaluation and Use of Automated Scoring*. Princeton, NJ, USA: Educational Testing Service, 2012. https://www.ets.org/Media/Research/pdf/RR-12-08.pdf
10. The Guardian, "Universities struggle with AI-generated essays," *The Guardian*, Jan. 16, 2023.
11. OECD, *Education at a Glance 2021*. Paris, France: OECD Publishing, 2021. https://www.oecd.org/education/education-at-a-glance/
12. UNESCO, *AI and Education: Guidance for Policy-makers*. Paris, France: UNESCO, 2021. https://unesdoc.unesco.org/ark:/48223/pf0000376709
13. C. O'Neil, *Weapons of Math Destruction*. New York, NY, USA: Crown Publishing Group, 2016. https://weaponsofmathdestructionbook.com/
14. European Commission, *Ethics Guidelines for Trustworthy Artificial Intelligence*. Brussels, Belgium: European Union, 2020. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
15. P. Liang, R. Bommasani, T. Lee, *et al.*, "Detecting AI-generated text," *arXiv preprint arXiv:2301.11305*, 2023. https://arxiv.org/abs/2301.11305