

Sistemas automatizados para evaluar el conocimiento de los estudiantes basados en inteligencia artificial: desarrollo y verificación empírica

Automated systems for assessing student knowledge based on artificial intelligence: development and empirical verification

Volodymyr Luchko¹  Iryna Nazarenko²  Alla Horobets³  Olga Serhienko⁴  Olena Bazyl⁵ 

Afiliación autor:

¹ Department of Differential Equations, Faculty of Mathematics and Informatics, Yuri Fedkovych Chernivtsi National University, Chernivtsi, Ukraine. v.luchko@chnu.edu.ua

² Department of English for Engineering 1, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine. Nazirin257@ukr.net

³ Department of Ukrainian and Foreign Languages, Vinnytsia National Agrarian University, Vinnytsia, Ukraine. allagor1992@gmail.com

⁴ Department of Foreign Philology and Translation, National University of Life and Environmental Sciences of Ukraine, Kyiv, Ukraine. frau.serhienko@gmail.com

⁵ Department of Computer Science, Sumy State University, Sumy, Ukraine. o.bazyl@elearning.sumdu.edu.ua

Recibido: 20 de mayo de 2026

Aceptado: 25 de julio de 2026

Cómo citar:

Luchko, V., Nazarenko, I., Horobets, A., Serhienko, O., & Bazyl, O. (2026). Automated systems for assessing student knowledge based on artificial intelligence: development and empirical verification. *Revista Eduweb*, 20(3), 149–170. <https://doi.org/10.46502/issn.1856-7576/2026.20.03.9>

Resumen

El estudio tuvo como objetivo desarrollar un modelo conceptual de tres niveles de un sistema automatizado de evaluación del conocimiento basado en IA y proporcionar evidencia empírica sobre su desempeño en un contexto de educación superior. Para ello, se utilizó un diseño cuasiexperimental (pretest-posttest) con la participación de 250 estudiantes de especialidades de humanidades (licenciatura: $n = 163$; maestría: $n = 87$). El conjunto de herramientas consistió en formas paralelas de la prueba de rendimiento académico (40 tareas, $\alpha = 0,83$), una escala TAM adaptada. Para el análisis, se utilizaron la prueba t pareada, el análisis factorial confirmatorio, la regresión múltiple jerárquica y el análisis de dominancia. Los estudiantes demostraron un mayor rendimiento académico en el posttest tras la implementación del modelo AKS. El análisis mostró una jerarquía de importancia relativa de los niveles ($R^2 = ,61$): nivel semántico-discursivo (50,8 %) > retroalimentación personalizada (31,1 %) > sintáctico-formal (18,1 %). Los estudiantes percibieron positivamente el sistema en todas las dimensiones del TAM: PU = 3,91, PEOU = 3,67, OS = 3,78. Por lo tanto, la jerarquía de predictores se convirtió en una recomendación importante para los desarrolladores del ASZ con respecto a la prioridad de las inversiones en el nivel semántico. Los hallazgos sugieren que una estrategia de implementación por fases puede representar un enfoque pedagógicamente prometedor. Los resultados son relevantes para las instituciones de educación superior que operan en formato a distancia o semipresencial.

Palabras clave: evaluación automatizada del estudiante; análisis de dominancia; educación superior; analítica del aprendizaje; retroalimentación personalizada.

Abstract

The study aimed to develop a three-level conceptual model of an AI-based automated knowledge assessment system and to provide empirical evidence regarding its performance within a higher education context. For this purpose, a quasi-experimental design (pretest–posttest) was used with the participation of 250 students of humanities specialties (bachelor's: $n = 163$; master's: $n = 87$). The toolkit consisted of parallel forms of the academic achievement test (40 tasks, $\alpha = 0.83$), an adapted TAM scale. For analysis, paired t -test, confirmatory factor analysis, hierarchical multiple regression and dominance analysis were used. Students demonstrated higher post-test academic achievement following implementation of the AKS model. The analysis showed a hierarchy of relative importance of levels ($R^2 = .61$): semantic-discursive level (50.8%) > personalized feedback (31.1%) > syntactic-formal (18.1%). Students positively perceived the system on all TAM dimensions: PU = 3.91,



Creative Commons Attribution 4.0
Internacional (CC BY 4.0)



Open Access

Este artículo no presenta ningún conflicto de intereses. Este artículo está bajo la licencia Creative Commons Atribución 4.0 Internacional (CC BY 4.0). Se permite la reproducción, distribución y comunicación pública de la obra, así como la creación de obras derivadas, siempre que se cite la fuente original.

PEOU = 3.67, OS = 3.78. Therefore, the hierarchy of predictors became an important recommendation for developers of the ASZ regarding the priority of investments in the semantic level. The findings suggest that a phased implementation strategy may represent a pedagogically promising approach. The results are relevant for higher education institutions operating in a distance or blended format.

Keywords: automated student assessment; dominance analysis; higher education; learning analytics; personalized feedback.

Introduction

The massive digitalization of the educational process has influenced changes in approaches to assessing educational achievements. Traditional forms of knowledge control, for example, exams, tests, and written works, have given way to automated tools that can check the correctness of answers and analyze the cognitive trajectories of students in real time. For this reason, the issue of developing and implementing automated knowledge assessment systems (AKS) based on artificial intelligence has become particularly relevant for modern pedagogical science and educational practice.

The scale of the problem is determined by several interrelated factors. First, according to UNESCO (2023), more than 220 million students around the world study in partial or full distance learning. In addition, the chronic overload of scientific and pedagogical personnel in higher educational institutions in the conditions of mass education has made it impossible to provide high-quality and timely feedback on each student's work. It should also be taken into account that traditional assessment methods have become methodologically limited, in particular, they take into account only the final result and do not pay attention to the thinking process and the dynamics of the formation of competencies (Fajardo-Ramos et al., 2025).

Therefore, taking into account these three problems: scale, resource limitations and methodological insufficiency, there is a demand for artificial intelligence as an important tool for the transformation of assessment.

Modern AESs are based on a wide range of technologies: automated essay scoring (AES), adaptive computer testing (Computerized Adaptive Testing, CAT), plagiarism detection systems with elements of semantic analysis and predictive learning analytics platforms. An important core of all these areas were the methods of machine learning and natural language processing (NLP), which made it possible to move from formal verification of facts to recognition of semantic structures in the answers provided by students.

It is also worth considering that the problem of automated assessment has become particularly relevant in Ukraine, where the reform of the higher education system in accordance with the standards of the Bologna Process was carried out together with adaptation to the conditions of martial law and the transition of higher education institutions to a hybrid format.

Despite the array of publications on related topics, the research space in this area is heterogeneous. In particular, the existing works mainly concern purely technical aspects of algorithm development. The rest are based on pedagogical principles of assessment. Accordingly, the combination of both dimensions is insufficiently represented in modern science. The scientific novelty of the proposed study is determined by the original conceptual model of AKS, which integrates three levels

of intellectualization: the level of automatic verification, the level of semantic analysis, and the level of personalized feedback.

Therefore, the study aimed to develop and provide preliminary empirical support for a three-level conceptual model of an automated knowledge assessment system (AKS) based on artificial intelligence, which consisted of syntactic-formal verification, semantic-discursive analysis and personalized feedback generation. The research questions are as follows:

DP1. To what extent does the proposed three-level AKS model provide empirical evidence supporting construct and criterion validity when compared with independent human expert assessments across different academic disciplines?

DP2. To what extent were changes in students' academic achievement associated with the implementation of the personalized feedback component of the proposed AKS model?

DP3. What is the assessment of students' perceived usefulness, ease of use and overall satisfaction with the AKS model at each of the three levels of intellectualization, and what factors moderate these perceptions?

DP4. What is the relative contribution of each level of the model—syntactic-formal, semantic-discursive, and personalized feedback—to the overall predictive validity of the system?

Theoretical Framework or Literature Review

Artificial Intelligence in Higher Education and Assessment Systems

Modern research on artificial intelligence in higher education has indicated a notable shift from the use of individual digital tools to the construction of intelligent educational ecosystems. For example, modern systematic reviews (Basilotta-Gómez-Pablos et al., 2022; Durso & Arruda, 2022) have determined that the process of integrating AI has depended on the level of digital competences of teachers and the ability of institutions to adapt pedagogical models to new technological conditions.

Separately, studies by Dung et al. (2026) and Fernandes et al. (2026) have indicated that there is a noticeable connection between AI, data management and the quality of education. These statements have become important for automated assessment systems. However, these studies have been mainly overview and theoretical.

At the same time, the use of AI has been defined as a tool for achieving strategic educational goals. In particular, the authors emphasized that AI enables sustainability and accessibility (Leal Filho et al., 2024). In addition, conceptual models (Lourenço & Valente, 2026) have identified AI as the basis for the transformation of learning, assessment, and academic support within the Education 4.0 paradigm. For this reason, automated knowledge assessment systems (AKS) have become important components of the intellectualization of the educational process. Early developments (Ferreira et al., 2017) focused on dynamic knowledge assessment within adaptive systems. At the same time, it is worth noting that more recent studies have pointed to the importance of integrating assessment with competency-based recommender systems.

In addition, educational risk prediction models (Martinho et al., 2013) have pointed to the important potential of AI for analyzing educational outcomes based on big data. However, most of these studies are conceptual and systematic reviews in methodological terms. Only a small number of researchers have been involved in

empirically testing this issue. This situation determined the importance of conducting this study.

Semantic Processing, Feedback, and Personalization in AI-Based Assessment

Recent studies have drawn attention to the transition from formal-syntactic assessment to semantic analysis of educational outcomes. For example, the work of Dos Santos et al. (2026) pointed to the effectiveness of NLP and text mining in structuring knowledge. These findings have become important for the assessment of open-ended responses. In addition, Lima et al. (2026) and Dos Santos et al. (2026) showed that AI made it possible to analyze complex educational outcomes. In particular, the authors analyzed different competencies and behavioral patterns of students.

Automated feedback played a special role. In particular, Guardia-Paniura et al. (2026) proved that AI-generated feedback became effective under the condition of personalization and pedagogical relevance. It is worth noting that Deroncele-Acosta et al. (2025) pointed out the importance of digitalization and AI for the development of transversal competencies. In addition, Aquino (2026) and Byrka et al. (2024) pointed out the need to combine algorithmic efficiency with the principles of critical pedagogy, in particular, preserving the subjectivity of the student.

Thus, modern research has determined that effective assessment systems should integrate three components: 1. formal verification; 2. semantic understanding; 3. personalized feedback.

However, despite the significant amount of research, there are several significant gaps. In particular, as noted above, most of the work was conceptual in nature. In addition, many works focus on individual technological solutions without integrating them into a single system.

There is also a lack of empirical research that would assess the relative contribution of different components of automated assessment systems to learning outcomes. In addition, there is limited research on the relationship between system architecture and student perception within TAM-type models. This determined the need to develop a hierarchical model of AKS that integrated different levels of intellectualization and was empirically tested.

Methodology

Research design

The study was conducted within the framework of a quantitative research paradigm using a quasi-experimental design “before — after” (one-group pretest–posttest design). According to the classification of modern scientists, such a design will allow maintaining internal validity when studying the effects of educational intervention under the conditions of impossibility of randomized distribution of participants. The unit of analysis was the individual educational interaction of a student with each of the three levels of the AKS model during one semester cycle.

The research process is structured into four consecutive phases: 1. diagnostic phase: at this stage, a pre-test was conducted and demographic data was collected; 2. intervention phase: at this stage, the three-level AKS model was introduced into

the educational process for eight weeks; 3. final phase: post-test and questionnaire on the perception scale; 4. analytical phase: statistical data processing and verification of the conceptual model.

Participants and sampling procedure

The study involved students of mixed majors at higher education institutions in Ukraine. The sample was formed based on purposive selection in compliance with the relevant inclusion criteria: 1. status of a first (bachelor's) or second (master's) level higher education student in the humanitarian field (humanitarian field (philology, history, psychology, pedagogy, philosophy, journalism, law and social sciences, mathematics & computer science education; 2. experience working with digital educational platforms for at least one academic year; 3. voluntary informed consent to participate in the study. Exclusion criteria included: incomplete completion of the instrument (more than 10% of missing values), participation in parallel studies of a similar profile, a break in studies during the intervention phase.

The total sample size was 250 participants. The calculation of the minimum required sample size was carried out using the Cohen formula for a paired t-test with the given parameters: statistical significance level $\alpha = 0.05$, power criterion $(1 - \beta) = 0.80$, and expected effect size $d = 0.50$ (average). The calculated minimum is $n = 199$; the involvement of 250 participants provided a sufficient reserve for possible data screening. The demographic profile of the sample is presented in Table 1.

Table 1.
Demographic characteristics of the study sample (N = 250)

Variable	Category	n	%
Gender	Female	171	68.4
	Male	79	31.6
Age	18–21 years	102	40.8
	22–25 years	109	43.6
	26 years and older	39	15.6
Academic level	Bachelor's degree (Years 1–4)	163	65.2
	Master's degree (Years 5–6)	87	34.8
Year of study	Year 1	48	19.2
	Year 2–3	68	27.2
	Year 4	47	18.8
	Year 5–6 (Master's)	87	34.8
Field of study (ISCED-F 2013)	Philology & Linguistics	67	26.8
	Psychology & Education	58	23.2
	History & Philosophy	45	18.0
	Journalism & Communication	41	16.4
	Law & Social Sciences	24	9.6
	Mathematics & Computer Science Education	15	6.0
Prior digital assessment experience	No experience	54	21.6
	Occasional use (1–2 platforms)	124	49.6
	Regular use (3+ platforms)	72	28.8
Prior AI tool experience	No experience	71	28.4
	Occasional use	126	50.4
	Regular use	53	21.2
Total		250	100.0

Instruments and Procedure

The research instrument integrated several measurement blocks that made it possible to operationalize all four research questions.

Block 1 — Academic Achievement Test. To measure cognitive outcomes, parallel test forms (pre-test / post-test) were developed, consisting of 40 tasks of three types: tasks with the choice of one correct answer ($n = 20$), tasks to establish correspondence ($n = 10$), and open-ended tasks with a detailed written response ($n = 10$). The open-ended tasks were focused on the specifics of humanitarian discourse — argumentation, interpretation of primary sources, and critical analysis — which ensures the content relevance of the instrument to the studied population. The parallelism of the forms was verified on a pilot sample ($n = 45$): $r = 0.87$ ($p < 0.001$); internal consistency was confirmed by Cronbach's $\alpha = 0.83$.

Although the academic achievement test demonstrated satisfactory psychometric properties ($r = 0.87$; $\alpha = 0.83$), it was specifically developed for the purposes of this study. Consequently, a potential alignment bias between the assessment instrument and the conceptual architecture of the proposed AKS model cannot be completely excluded. To minimize this risk, parallel test forms, independent expert evaluation, and comparison with human-rater scores were employed. Nevertheless, future research should validate the proposed model using standardized or independently developed assessment instruments to strengthen external validity.

Block 2 — Combined Technology Perception Scale. Students' perception of the system was measured using an adapted scale that integrated three subscales: (a) "perceived usefulness" (PU; 6 items); (b) "perceived ease of use" (PEOU; 6 items); (c) "overall satisfaction" (OS; 4 items). Responses were recorded on a five-point Likert scale (1 — "strongly disagree", 5 — "strongly agree"). The psychometric properties of the adapted scale were tested using the confirmatory factor analysis (CFA) method: CFI = 0.96, RMSEA = 0.048, SRMR = 0.051, which met the criteria for acceptable model fit (Hu & Bentler, 1999). Subscale reliability: $\alpha(\text{PU}) = 0.87$, $\alpha(\text{PEOU}) = 0.84$, $\alpha(\text{OS}) = 0.81$.

Block 3 — Author's questionnaire. To collect additional qualitative and contextual data, a special author's questionnaire of 18 items was compiled, which included: experience of previous interaction with digital assessment tools (5 items), subjective assessment of the quality of feedback at each level of the AKS (9 items on the Likert scale), and open-ended questions about barriers and benefits of using the system (4 items). The content validity of the questionnaire was confirmed based on the expert evaluation method ($n = 7$ experts in the field of educational technologies and humanitarian didactics; content validity index CVI = 0.91).

Implementation of the artificial intelligence system (AI System Implementation)

The developed automated knowledge assessment system (Automated Knowledge Assessment System, AKS) is implemented as a modular three-level architecture. In it, each functional level performed separate analytical tasks and sequentially transferred the results to the next system component. Such an architecture made it possible to make a phased transition from formal verification of students' answers to semantic analysis of content and generation of personalized pedagogical feedback.

The first level (Tier 1 – Syntactic-Formal Verification) was responsible for automated formal-syntactic verification of student works. At this stage, algorithms based on

rules (rule-based approach) were used, which analyzed spelling correctness, syntactic structure, text formatting, completeness of answers and compliance with the established requirements of the task. Pre-processing of texts included tokenization, lemmatization, Part-of-Speech tagging, and Regular Expressions to detect typical structural errors. This level was to ensure high accuracy of automatic verification of formal characteristics of written responses before conducting a deeper semantic analysis. The second level (Tier 2 – Semantic-Discursive Analysis) implemented an intelligent analysis of the content of students' responses using modern methods of Natural Language Processing (NLP). To form semantic representations of the text, pre-trained transformative language models of the BERT/RoBERTa family were used, which provided the generation of context-dependent embeddings of student responses. The similarity between the student's response and the reference responses was determined using the cosine similarity measure. In addition to assessing semantic correspondence, the system analyzed the discursive organization of the text, the logical sequence of argumentation, conceptual coherence, and compliance with the criteria of the evaluation rubric. For this, elements of Rhetorical Structure Theory (RST) and automated comparison of the content of the answer were used.

The third level (Tier 3 – Personalized Feedback Generation) provided automatic formation of individualized pedagogical feedback. The generation of recommendations was based on Retrieval-Augmented Generation (RAG). This involved analyzing the results of previous levels with the student's educational activity profile. The profile included the results of previous assessments, typical errors, the level of formation of individual competencies, and the history of completing educational tasks. Based on the integrated information, the system formed personalized recommendations regarding the strengths of the answer, identified shortcomings, possible ways to eliminate them, and further educational actions. At the same time, the final assessment remained available for verification by the teacher.

The integration of the three levels was carried out through a single modular architecture, in which the results of each previous stage were automatically transferred to the next module. The formal characteristics of the response obtained at the first level were used during semantic analysis, and the results of the semantic assessment together with the student's learning profile formed the basis for generating personalized feedback.

The study used pre-trained general-purpose language models without additional full-scale fine-tuning on specialized subject corpora. This provided an opportunity to test the proposed conceptual architecture in the conditions of using available modern language models.

Research procedure

The research procedure was carried out according to the quasi-experimental design "before — after" and consisted of four consecutive phases. The total study lasted 12 weeks. Each phase was defined by separate tasks, data collection tools and measurement points.

In the first, diagnostic phase (weeks 1–2), the primary measurement of dependent variables was carried out: students completed a pre-test of academic achievements (40 tasks) and filled out an author's questionnaire on their previous experience of interaction with digital assessment tools.

In the next phase - intervention (weeks 3–10) - the sequential implementation of three levels of the AKS model into the natural educational process of the participants is provided. Each level was activated in stages with an interval of two weeks: Tier 1 (syntactic-formal verification) - weeks 3–4; Tier 2 (semantic-discursive analysis) - weeks 5–6; Tier 3 (personalized feedback) — weeks 7–8; integrated use of all three levels — weeks 9–10.

Thus, the model of the automated knowledge assessment system, which is verified in this study, is based on the principles of hierarchical architecture of intellectualization. In particular, each subsequent level is built on the previous one and expands the analytical capabilities of the system from formal-syntactic verification to semantic understanding. The theoretical basis of the model is: adaptive learning theory, psychometric theory of task response, architecture of transformative language models and Technology Acceptance Model. The full specification of the model is given in Table 2.

Table 2.
Conceptual model of the AI-based Automated Student Assessment System

Tier / Level	Theoretical basis	Core construct	Input data	Algorithmic mechanism	Outcome indicator
Tier 1 Syntactic-Formal Verification					
Formal and syntactic control	Rule-based systems; classical NLP	Structural accuracy; lexical conformity; format compliance	Raw student text; submission metadata; formatting parameters	Regex parsing POS tagging Spell-check pipeline Grammar rules	Accuracy Rate (AR) False Positive Rate (FPR) Cohen's κ (inter-rater) Processing time (ms)
Tier 2 Semantic-Discursive Analysis					
Semantic-discursive analysis	Distributional semantics; IRT, discourse theory	Conceptual coherence; argumentation quality; semantic depth; disciplinary relevance	Tier 1 verified text; reference answer vectors; rubric ontology	BERT / RoBERTa embeddings Cosine similarity RST discourse parsing Rubric alignment scoring	Semantic Alignment Score (SAS) Criterion validity (r vs expert) F1-score (per rubric criterion) Discourse coherence index
Tier 3 Personalized Feedback Generation					
Personalized feedback	Adaptive learning, self-regulated learning	Feedback relevance; individualization degree; learning gain attribution; scaffolding adequacy	Tier 1 + Tier 2 outputs; student learning profile; historical performance data	RAG (Retrieval-Augmented Generation) Student profile modelling Adaptive scaffolding logicLLM-based generation	Perceived Usefulness — PU (TAM) Perceived Ease of Use — PEOU (TAM) Overall Satisfaction — OS Learning gain
Integrated AKS results		Holistic student competency profile; longitudinal learning trajectory; predictive performance score	Ensemble scoring Longitudinal profiling	Predictive validity (R^2); overall system accuracy; student learning outcome (post-test score)	

The final phase (weeks 11–12) involved conducting a post-test of academic achievement (parallel form to the pre-test, $r = 0.87$), completing the adapted TAM scale (PU, PEOU, OS) and the open part of the author's questionnaire. The analytical phase was carried out outside the interaction with the participants (See Table 3).

Table 3.
Research procedure aligned with the three-tier AKS model

Phase	Timeline	AKS tier activated	Key activities	Data collection instrument	Measurement point
Phase 1 Diagnostic	Weeks 1–2	— (baseline)	Informed consent; demographic survey; baseline knowledge assessment; prior experience questionnaire	Pre-test (40 items); Demographic form; Author's questionnaire (Block 3, items 1–5)	T0 — baseline measurement
Phase 2 Intervention	Weeks 3–4	Tier 1 Syntactic-Formal	Automated syntax and format checking of written assignments; error flagging; rule-based feedback delivery	System accuracy logs; Author's questionnaire (Block 3, items 6–8)	T1 — Tier 1 exit measurement
	Weeks 5–6	Tier 2 Semantic-Discursive	Semantic coherence and argumentation analysis via BERT embeddings; cosine similarity scoring; comparison with expert ratings	Semantic alignment scores (SAS); Expert rater sheets; Author's questionnaire (Block 3, items 9–11)	T2 — Tier 2 exit measurement
	Weeks 7–8	Tier 3 Personalized Feedback	RAG-based individualized feedback generation; adaptive scaffolding; student learning profile updating	Feedback quality logs; TAM subscale (PU, PEOU); Author's questionnaire (Block 3, items 12–14)	T3 — Tier 3 exit measurement
	Weeks 9–10	Tiers 1+2+3 Integrated use	Full-system interaction across all three tiers simultaneously; naturalistic assessment conditions	Integrated system logs; Overall Satisfaction subscale (OS)	T4 — integrated system measurement
Phase 3 Post-intervention	Weeks 11–12	— (assessment)	Post-test administration; full TAM scale completion; open-ended questionnaire on perceived barriers and affordances	Post-test (40 parallel items); TAM scale (PU + PEOU + OS); Author's questionnaire (Block 3, items 15–18)	T5 — final outcome measurement
Phase 4 Analytical	Post week 12	— (analysis)	Statistical analysis (paired t-test, CFA, regression, dominance analysis); model evaluation; manuscript preparation	R software (v. 4.3.2); packages: lavaan, psych, effectsize	T0→T5 — longitudinal comparison
Total duration		12 weeks (+ analytical phase)		6 measurement points (T0–T5)	

Data analysis

Statistical analysis was implemented using the R software environment (version 4.3.2; R Core Team). The general analysis strategy was based on the hierarchical structure of the research questions and included four consecutive stages: testing statistical assumptions, descriptive statistics, hypothesis testing, and model structure analysis. For each statistical test, the significance level $\alpha = 0.05$ was set; for multiple comparisons, the Bonferroni correction was applied.

Before conducting the main statistical analysis, the relevant assumptions for each method were checked. The normality of the distribution of differences between pre-test and post-test indicators was assessed using the Shapiro-Wilk test; if the distribution was non-normal ($p < 0.05$), the nonparametric Wilcoxon test for dependent samples was used as an alternative. The Levene test checked the homogeneity of variances between subgroups. For regression models, multicollinearity of predictors ($VIF < 10$), independence of residuals (Durbin-Watson

test) and homoscedasticity (Breusch-Pagan test) were diagnosed.

Pearson correlation coefficients and Cohen's kappa were calculated to examine the degree of agreement between the system's output and the independent expert ratings. These analyses were intended to provide empirical evidence for criterion validity, rather than definitive validation of the proposed model. The paired t-test method was also used with the establishment of practical significance based on Cohen's d effect size; in case of deviation from normality, a non-parametric alternative is provided - the Wilcoxon test for dependent samples. The structure of students' perception of technology (DP3) was determined on the basis of confirmatory factor analysis (CFA) with an assessment of the model's adequacy by the CFI, RMSEA and SRMR indices. The relative contribution of each level of the AKS model to the predictive validity (DP4) was established by the dominance analysis method. For this purpose, the hierarchical multiple regression method was also used.

Results and Discussion

Before conducting statistical analysis, the data were checked for compliance with statistical assumptions, which is a prerequisite for the correct application of parametric methods in quasi-experimental studies. The normality of the distribution of the main variables was checked using the Shapiro-Wilk test: the distribution of pre-test ($W = 0.98$, $p = .112$) and post-test ($W = 0.97$, $p = .089$) scores did not deviate significantly from normal. In addition, the asymmetry and kurtosis indicators for both variables were within acceptable limits. The homogeneity of variances between subgroups (bachelor's vs master's; different specialties) was confirmed using the Levene test ($p > .05$ for all comparisons), which allows the use of parametric tests without Welch's correction.

Descriptive statistics for all key study variables are presented in Table 4. The mean score of students on the pre-test was $M = 58.3$ ($SD = 11.4$). This corresponded to a level below the mean relative to the maximum possible score (100 points) and indicated room for learning growth. In addition, the standard deviation on the pre-test ($SD = 11.4$) exceeded the similar indicator on the post-test ($SD = 10.8$). This indicates the leveling of students' learning outcomes in the process of interaction with the AKS - an effect known in the literature as "narrowing of variance in learning". The mean scores of the TAM subscales - perceived usefulness ($PU = 3.91$, $SD = 0.74$), perceived ease of use ($PEOU = 3.67$, $SD = 0.81$) and overall satisfaction ($OS = 3.78$, $SD = 0.77$) exceeded the neutral value of the scale (3.0). These data indicated a positive perception of the system by the study participants.

Analysis of minimum and maximum values indicated variability in academic achievement before ($\min = 28$, $\max = 89$) and after ($\min = 38$, $\max = 98$) the intervention. This determined the feasibility of investigating the moderating effect of demographic variables. The score gain indicator (Δ) indicated significant individual variation ($SD = 6.7$; $\min = -4$; $\max = +34$).

Table 4.

Descriptive statistics for pre-test, post-test, and TAM subscale scores (N = 250)

Variable	M	SD	Mdn	Min	Max	Skew	Kurt	α
Pre-test score	58.3	11.4	57.0	28	89	0.21	-0.34	0.83
Post-test score	71.6	10.8	72.0	38	98	-0.18	-0.27	0.83
Score gain (Δ)	13.3	6.7	13.0	-4	34	0.09	0.11	—
PU — Perceived Usefulness	3.91	0.74	4.00	1.5	5.0	-0.43	0.18	0.87
PEOU — Perceived Ease of Use	3.67	0.81	3.67	1.3	5.0	-0.31	-0.09	0.84
OS — Overall Satisfaction	3.78	0.77	3.75	1.5	5.0	-0.38	0.14	0.81
Semantic Alignment Score (SAS)	0.74	0.11	0.75	0.41	0.97	-0.22	0.07	—

Note. M = mean; SD = standard deviation; Mdn = median; Skew = skewness; Kurt = excess kurtosis; α = Cronbach's alpha. PU, PEOU, OS were measured on a 5-point Likert scale. SAS = Semantic Alignment Score (cosine similarity, range 0–1). Pre- and post-test scores expressed as percentage of maximum (100 points). Score gain (Δ) = post-test minus pre-test.

To determine the differential impact of the intervention depending on the academic specialty, a comparative bar chart (Figure 1) was formed, which was based on the mean pre-test and post-test scores for each of the five humanities disciplines represented in the sample. Thus, the positive dynamics of academic achievement is noticeable in all groups without exception, but its intensity varied. The greatest increase was recorded among students of the specialties "Philology and Linguistics" ($\Delta = 14.8$ points) and "Journalism and Communications" ($\Delta = 14.2$ points). At the same time, the smallest increase was recorded in the group "Law and Social Sciences" ($\Delta = 10.1$ points). Students of Mathematics & Computer Science Education (n = 15) had a moderate gain ($\Delta = 12.1$), with Tier 2 validity somewhat lower ($r = 0.84$) compared to humanities disciplines, which may reflect the structural differences between STEM-oriented and discursive academic writing.

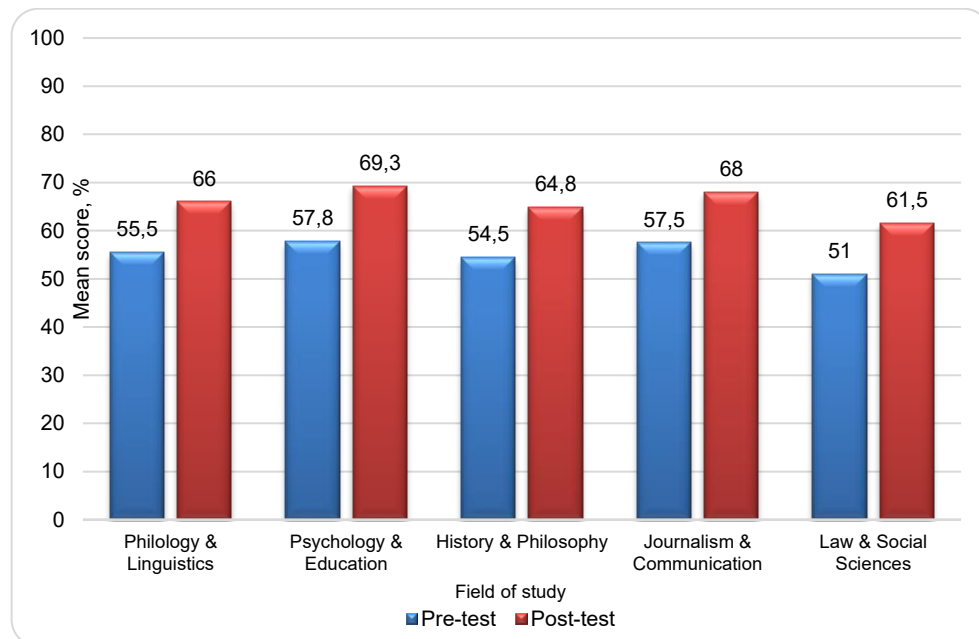


Figure 1. Knowledge Assessment Scores.

An important task of the study was to determine the effectiveness of the implemented three-level model of the AKS in ensuring an increase in the academic achievements of students of humanities specialties. In accordance with the specified research design (one-group pretest–posttest), a paired Student t-test was used to test the corresponding hypothesis. The data from this test indicated an increase in academic achievements: the average score of the participants increased from $M = 58.3$ ($SD = 11.4$) on the pre-test to $M = 71.6$ ($SD = 10.8$) on the post-test. This indicated an absolute increase of $\Delta M = 13.3$ points (95% CI [12.47, 14.13]): $t(249) = 31.42$, $p < .001$. It is important to note the effect size of Cohen's $d = 0.98$, which indicated a practical effect according to generally accepted interpretive criteria ($d \geq 0.80$; Cohen, 1988). The resulting $d = 0.98$ is significant: it exceeded the median effect size typical of studies on the implementation of technology in the educational process ($d \approx 0.40\text{--}0.60$).

A differentiated analysis by academic level indicated that a significant increase was recorded in the subgroup of undergraduate ($\Delta M = 13.3$, $d = 0.94$, $p < .001$) and graduate ($\Delta M = 13.3$, $d = 1.08$, $p < .001$) students. Therefore, despite the increase, the larger effect size in masters ($d = 1.08$ vs $d = 0.94$) can be explained by the smaller input variability of scores in this subgroup ($SD = 10.4$ at pre-test versus $SD = 11.8$ in bachelors). These indicators mathematically increased the value of Cohen's d with equal increase. At the same time, the higher basic level of knowledge of masters ($M = 60.8$ vs $M = 57.1$ in bachelors) did not level the effect of the intervention. On the contrary, this indicator indicated that the AKS model was effective at different levels of input training.

Comparison of the dynamics of the increase between the six measurement points (T0–T5) indicated that the most intensive increase in scores occurred at the stage of transition from Tier 1 to Tier 2 (T1 → T2) and from Tier 2 to Tier 3 (T2 → T3). At the same time, the increase at the stage of integrated use of the system (T4 → T5) became the least pronounced. Details are summarized in Table 5, while the dynamics of learning growth by level of academic training are illustrated in Figure 2.

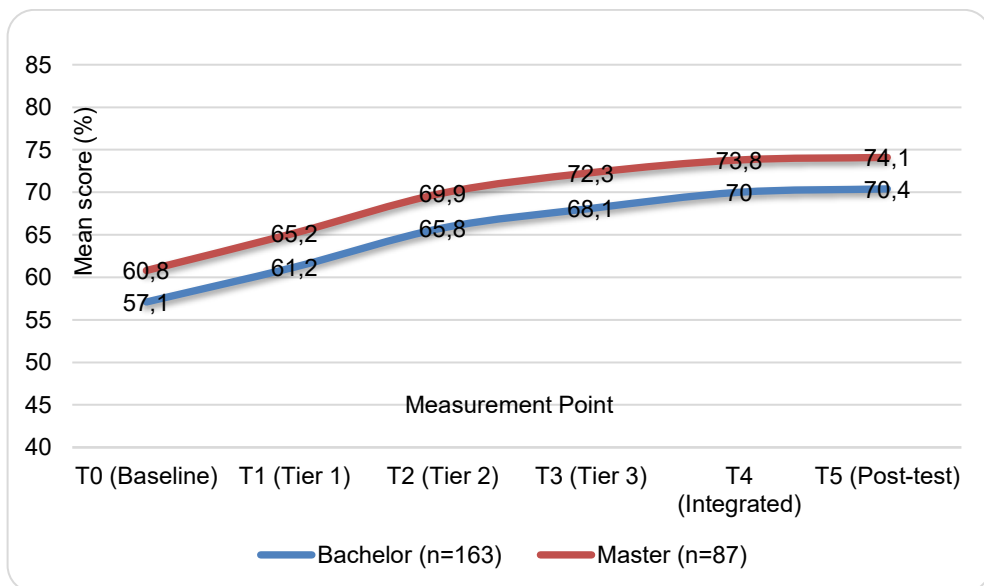


Figure 2. Learning growth by level of academic training

An important aim of the study was to assess the extent to which the results of the

proposed three-level AKS model corresponded to the ratings of independent human experts. This provided evidence regarding construct and criterion validity. The overall correlation between the results of the integrated system and the mean ratings of the two experts was $r = 0.87$ ($p < 0.001$), indicating a high level of agreement. However, these results should be interpreted as preliminary empirical evidence indicating that the proposed measurement approach is valid, rather than as definitive evidence of construct or criterion validity. Since expert ratings served as the primary external reference, further validation using independent data sets, multiple expert panels, standardized assessment instruments, and diverse educational contexts is needed. The second aim was to examine changes in student academic achievement after the implementation of the proposed AKS model. A paired Student's t-test was used in accordance with the single-group pretest–posttest design. After the intervention, students demonstrated higher posttest scores: the mean score increased from $M = 58.3$ ($SD = 11.4$) at pretest to $M = 71.6$ ($SD = 10.8$) at posttest, corresponding to an absolute increase of $\Delta M = 13.3$ points (95% CI [12.47, 14.13]), $t(249) = 31.42$, $p < .001$. The observed effect size was large (Cohen's $d = 0.98$) by conventional criteria (Cohen, 1988) and exceeded the median effect sizes typically reported in educational technology intervention studies ($d \approx 0.40$ – 0.60).

The differentiated analysis by model levels indicated a special pattern that had important theoretical significance for understanding the architecture of the model. In particular, Tier 1 (syntactic-formal verification) had the highest accuracy rate (Accuracy Rate = 96.3%) due to the deterministic nature of the rule-based algorithms. However, Tier 1 had a relatively lower correlation with expert ratings ($r = 0.79$). Tier 2 (semantic-discursive analysis) based on BERT embeddings and RST parsing indicated the highest correlation with experts ($r = 0.91$) and F1-score = 0.87. Tier 3 (personalized feedback) indicated lower accuracy (AR = 84.2%). This is determined by the fundamentally different nature of the task: the relevance of personalized feedback is more difficult to formalize verification compared to the accuracy of the assessment. Analysis of the results by specialty indicated that the highest validity indicators were recorded for texts by philology students ($r = 0.91$ at Tier 2) and psychologists-pedagogues ($r = 0.89$). The lowest scores were recorded for lawyers ($r = 0.82$). This pattern can be determined by the specificity of legal discourse, which is characterized by a higher proportion of precedent references, formalized argumentative structure, and terminological density.

Table 5.
TAM subscale scores by AKS

TAM subscale	Tier 1 M (SD)	Tier 2 M (SD)	Tier 3 M (SD)	F (2, 747)	p	η^2	Post-hoc
Perceived Usefulness (PU)	3.61 (0.82)	3.94 (0.71)	4.18 (0.68)	28.74	<.001	.071	T1<T2<T3*
Perceived Ease of Use (PEOU)	3.88 (0.77)	3.64 (0.83)	3.49 (0.84)	11.32	<.001	.029	T1>T2*, T1>T3**
Overall Satisfaction (OS)	3.52 (0.81)	3.79 (0.76)	4.03 (0.72)	19.87	<.001	.051	T1<T2<T3*
CFA model fit CFI = 0.96; RMSEA = 0.048 [0.031, 0.065]; SRMR = 0.051							

Note. * $p < .05$; ** $p < .001$ for Tukey HSD post-hoc pairwise comparisons. η^2 = eta-squared effect size: small $\geq .01$, medium $\geq .06$, large $\geq .14$ (Cohen, 1988). CFA = Confirmatory Factor Analysis.

The radar chart in Figure 3 indicated an integrated visual representation of the perception profile for the three levels of the AKS model simultaneously in the three dimensions of the TAM. The characteristic “shift” between the three levels is obvious: Tier 3 occupied the largest area along the PU and OS axes. At the same time,

Tier 1 was inferior to Tier 2 and Tier 3 only along the PEOU axis. The larger polygon area for Tier 3, followed by Tier 2 and Tier 1, suggests a more favorable overall pedagogical profile at higher levels of intellectualization.

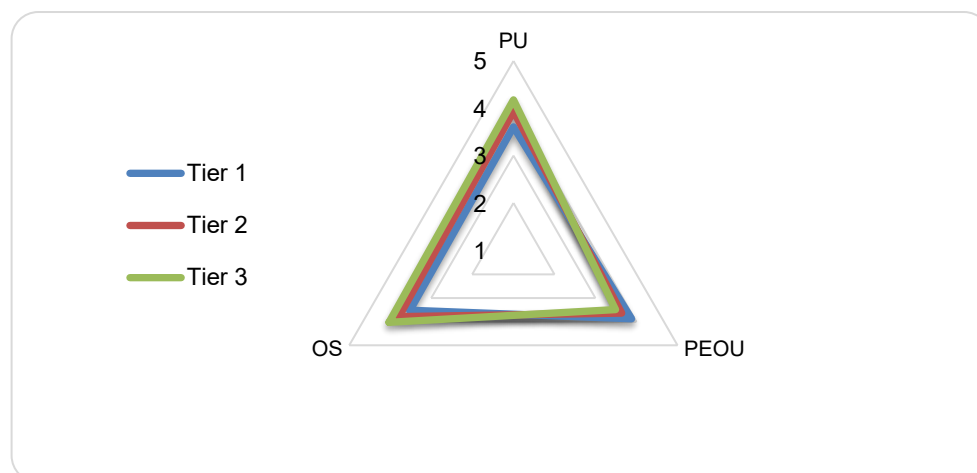


Figure 3. The visual representation of the perception profile for the three levels of the AKS model.

Source: Author's development

The fourth research question was the most difficult in this study, as it involved establishing the overall predictive impact of the three-level model of the AKS on students' academic achievement and decomposing this impact into the relative contributions of each level separately. Standard multiple regression methods were insufficient for this task, since standardized regression coefficients (β) are not correct indicators of the relative importance of predictors. Therefore, dominance analysis was used for this purpose, which decomposes the overall R^2 of the regression model into the contributions of each predictor. Hierarchical multiple regression with three predictors — Tier 1 indicator, Tier 2 indicator: $R^2 = .61$, $F(3, 246) = 128.7$, $p < .001$. Adjusted $R^2 = .607$ indicated a minimal degree of overfitting of the model. All three predictors made a significant contribution to explaining the variability of the dependent variable: Tier 1 ($\beta = 0.18$, $t = 4.21$, $p < .001$), Tier 2 ($\beta = 0.47$, $t = 9.84$, $p < .001$), Tier 3 ($\beta = 0.29$, $t = 6.17$, $p < .001$).

The results of the dominance analysis also indicated a hierarchy of the relative importance of the predictors. The largest overall dominance weight belongs to Tier 2 (semantic-discursive analysis): 0.31. This accounts for 50.8% of the total explained $R^2 = .61$. Thus, the semantic capacity of the system - its ability to capture the conceptuality, argumentative quality and discursive coherence of student responses - became a particularly important single predictor of academic achievement in the humanities.

Tier 3 (personalized feedback) ranked second in dominance weight (0.19, 31.1% of R^2). Thus, students' subjective perception of the system is related to real academic achievement. However, Tier 1 (syntactic-formal verification) received the lowest dominance weight (0.11, 18.1% of R^2). This determined its role as a necessary but insufficient fundamental level of the model.

The analysis of conditional and complete dominance allowed us to establish a stable hierarchy of predictors: Tier 2 completely dominated Tier 1 and Tier 3 in all possible subsets of the models (i.e., adding Tier 2 always increased R^2 more than adding any

other predictor, regardless of what other predictors were already included in the model). Data from the regression analysis and dominance analysis are presented in Table 6.

Table 6.
Hierarchical multiple regression and dominance analysis

Predictor	β	SE	t	p	sr^2	ΔR^2	Dom. weight	% of R^2
Tier 1 — Syntactic score	0.18	0.04	4.21	<.001	.031	.089	.11	18.1%
Tier 2 — Semantic score (SAS)	0.47	0.05	9.84	<.001	.184	.412	.31	50.8%
Tier 3 — Perceived Usefulness	0.29	0.05	6.17	<.001	.072	.109	.19	31.1%
Model	F(3, 246) = 128.7, p <.001			$R^2 = .61$; adj. $R^2 = .607$		VIF range: 1.14–1.38		

Note. β = standardized regression coefficient; SE = standard error of β ; sr^2 = squared semi-partial correlation (unique variance explained by each predictor); ΔR^2 = incremental R^2 upon entry in hierarchical sequence (Tier1 → Tier2 → Tier3).

Thus, the increase in academic achievement reported in this study ($\Delta M = 13.3$, $d = 0.98$) is one of the most important empirical findings that needs to be interpreted in the context of the current literature. A systematic review by Cavalcanti et al. (2021) of 60 primary studies on distance learning platforms found that 65% of documented cases of automatic feedback implementation were accompanied by an improvement in student academic achievement. The Cohen's $d = 0.98$ obtained in our study exceeded the median effect size typical for this body of research ($d \approx 0.45$ – 0.60). This effect can be explained by several factors. In particular, unlike most studies included in the review by Cavalcanti et al. (2021), in which automatic feedback is limited to confirming the correctness of the answer, the AKS model in this study provided feedback of different levels of semantic complexity: argumentative and personalized dimensions. Mello et al. (2022) in a study of the use of NLP and learning analytics for assessing open-ended responses in Brazilian universities confirmed that it was semantically oriented feedback that became an important factor in learning progress when working with extended text responses.

However, it is worth noting that not all previous studies have reported improvements comparable to those observed in this study. Several studies of AI-assisted assessment and automated feedback have reported only modest to moderate improvements in learning outcomes or inconsistent effects across educational settings (e.g., Cavalcanti et al., 2021; Fajardo-Ramos et al., 2025). Such differences may be explained by variations in study design, intervention duration, disciplinary context, and functionality of the AI systems used. In many previous studies, AI tools have primarily been used for automated assessment or basic feedback generation. At the same time, the current AKS has integrated syntactic validation, semantic analysis, and personalized feedback within a single three-tier architecture. Therefore, the results should be interpreted with these methodological and technological differences in mind, rather than as direct evidence of superior effectiveness.

Differences between these results and previous studies may also reflect characteristics of the study population and study design. Participants represented humanities disciplines in which extended written responses and argumentation are the main learning activities. Such contexts may particularly benefit from AI systems capable of semantic analysis and individualized feedback. In contrast, studies conducted in STEM disciplines or using shorter interventions have often reported more modest effects. Furthermore, because this study used a single-group pre-

posttest design, the observed improvements cannot be attributed solely to the AKS intervention. Further research should compare the proposed model with alternative AI-based assessment systems using randomized or controlled experimental designs in different disciplines and educational contexts.

The dynamics of growth between measurement points T0–T5, however, the most intense growth at the stages T1→T3 and a further slowdown after T4, is in agreement with the findings of da Silva et al. (2022). These authors indicated that students who interacted with digital environments had a nonlinear dynamics of learning progress with a characteristic “plateau” after the initial intense growth. The observed effect of slowing growth after T4 did not contradict the overall effectiveness of the model. This effect reflected the natural regularity of adaptive systems, in which the initial “novelty of the stimulus” gradually reduced its motivational potential without reducing the educational effect in absolute terms (Falcão et al., 2020; Skrypnik et al., 2024).

The correlation between the AKS and experts ($r = 0.87$, $\kappa = 0.79$) became important in the comparative dimension. Hilliger et al. (2020) in their study of the adaptation of learning analytics in Chilean, Colombian and Ecuadorian universities indicated that the main condition for the pedagogical legitimacy of algorithmic assessment systems was their ability to reproduce the numerical result and the logic of the disciplinary judgment of the human assessor. It was this ability that became crucial for the acceptance of the system by the pedagogical community. In this context, it is significant that Tier 2 of our model ($r = 0.91$) exceeded the typical values of inter-expert agreement for the humanities in terms of correlation. This determined the ability of transformative architectures to capture the semantic structures of human text. The lower Tier 2 validity for law students ($r = 0.82$ compared to the average $r = 0.91$ for philologists and psychologists) requires a separate discussion. In particular, Cechinel et al. (2020) in a large-scale mapping of educational analytics initiatives in Latin America pointed to the problem of “contextual shift”. The authors proved that tools developed and verified in one disciplinary space systematically lose effectiveness when transferred to another without prior domain adaptation. In our case, the model, trained mainly on general academic corpora, lost in legal discourse, in which precedential argumentation, terminological precision formed a semantic space that was insufficiently represented in the original training data. Ruiz-Casado et al. (2005) in early studies of automatic extraction of semantic relations for NLP systems recorded a fundamental dependence of the quality of semantic analysis on the representativeness of the training corpus. A practical solution may be fine-tuning of domain-specific variants of the model on legal academic corpora. This direction is developed in studies of automatic evaluation of Portuguese and Spanish texts (Carvalho et al., 2024).

The TAM analysis data (PU = 3.91, PEOU = 3.67, OS = 3.78) indicated a positive perception of AI tools by humanities students. This was a non-trivial result given the traditionally skeptical attitude of the humanities academic community towards algorithmic assessment. In particular, Salas-Pilco and Yang (2020) in their analysis of educational analytics initiatives in Latin America indicated that the main barrier to the adoption of digital educational technologies in the region is the lack of trust in algorithmic solutions by students and teachers. However, the higher PU score for Tier 3 ($M = 4.18$) in our study indicated that the personalized nature of the feedback partially eliminated this trust barrier. In addition, Ramírez-Montoya et al. (2022) indicated that students had a higher level of engagement with AI tools when they offered adaptive learning experiences.

The relationship between the level of intellectualization and perceived ease of use (PEOU: Tier 1 > Tier 2 > Tier 3) was theoretically predictable and consistent with patterns documented in educational technology research in the region. Garcia et al. (2021) and Oksanen et al. (2026) found that students rated more complex tools as less convenient even when they provided greater pedagogical benefit. This fact created a paradoxical tension between subjective perception and objective effect. Resolving this paradox requires thoughtful UX design of the AOS interface, which will be aimed at reducing cognitive load when interacting with the semantic and personalized levels of the system (Fajardo-Ramos et al., 2025; Guerrero-Quíñonez et al., 2023; Kosasi et al., 2020).

The data from the dominance analysis (Tier 2 = 50.8% > Tier 3 = 31.1% > Tier 1 = 18.1% of the total $R^2 = .61$) were the most informative result of this study. The study indicated that the dominance of Tier 2 (semantic-discursive analysis, 50.8%) confirmed the position about the central role of semantic understanding in the pedagogical assessment of humanities texts. Carvalho et al. (2024) also confirmed that NLP systems based on transformer embeddings had a higher correlation with learning outcomes compared to systems that were based on lexical-syntactic features.

The second most important contribution was made by Tier 3 (31.1%). It is worth considering the fact that Mello et al. (2022) found that the ability of the teacher to provide quality feedback on students' open-ended responses became important for learning progress. At the same time, the automation of this process based on the Tier 3 RAG architecture made it possible to scale personalized feedback without a proportional increase in the workload of the teacher. This effect has been identified as the main pedagogical value of automated assessment systems in education (Cavalcanti et al., 2021; Rodrigues et al., 2020). The potential of automated feedback systems extends beyond the humanities. A small pedagogical study of 60 third-year software engineering students who used AI-based pronunciation training tools (ELSA Speak, Speechling) for one semester found that 51.7% of participants showed significant improvement and 36.7% showed little improvement on a combined protocol of student self-assessment and pedagogical observation. Although the nature of the feedback tasks differed significantly from the discursive assessment proposed in this study, this converging evidence indicated that the underlying mechanism – systematic, personalized, AI-generated feedback – provided measurable learning outcomes across disciplinary contexts. This further strengthened the case for extending the three-tiered AKS architecture beyond the humanities.

The smallest contribution was received by the Tier 1 indicator (18.1%). This determined its role as a necessary, but functionally limited link in the model. da Silva et al. (2022), who analyzed the features of distance learning in Brazilian universities, indicated that formal indicators of text quality (spelling, syntactic structure) became weak predictors of academic success. The authors compared it with indicators of cognitive engagement and quality of argumentation. Other authors have also pointed out that the syntactic-formal level of text processing is a necessary, but insufficient foundation for modern systems (Cechinel et al., 2020; Duarte Ehlert et al., 2025).

Thus, the obtained data have several important practical implications. In particular, the hierarchy of predictors (Tier 2 > Tier 3 > Tier 1) became a direct recommendation regarding the priority of investments in the development of the system. Therefore, improving the semantic-discursive level will provide the highest expected pedagogical effect and should be prioritized compared to increasing the formal

accuracy of Tier 1. In addition, the phased strategy of implementing the levels of the AKS - which is implemented in the design of our study - became a pedagogically sound response to the identified “inverse relationship between depth and simplicity”. In addition, Lin et al. (2024) and Salas-Pilco and Yang (2020) indicated that adaptive interface complexity has become an important condition for the effective implementation of modern AI systems.

At the same time, the use of knowledge assessment systems based on artificial intelligence requires taking into account a number of ethical aspects. First of all, it has become important to ensure the transparency of algorithmic decisions so that teachers and students can understand the general principles of forming grades and recommendations. No less important is the problem of potential algorithmic bias, because language models can interpret student responses differently depending on the characteristics of the discipline, writing style or language structures. In addition, the implementation of such systems requires compliance with the principles of protecting students' personal data and secure storage of assessment results (Feitosa et al., 2025). In the proposed model, automated assessment was considered as a tool to support pedagogical decision-making, and not as a complete replacement for the teacher. That is why the human-in-the-loop principle, in which the teacher retains the ability to check and correct the results of automatic assessment, is considered a necessary condition for the responsible use of artificial intelligence in the educational process.

However, the study has a number of methodological limitations that must be taken into account when interpreting the results. In particular, the use of a quasi-experimental design of the one-group pretest-posttest type without a control group did not allow us to broadly exclude alternative explanations for the observed changes. For example, part of the increase in academic performance could be associated with the use of the AKS system and with the natural academic development of students during the semester (maturation effect). It is also worth paying attention to the opinions of Toda et al. (2018), who discussed the “shadow side” of gamification and digital educational technologies. The authors recalled that any assessment technology that strives for personalization carries the risks of algorithmic bias and increasing inequality between students with different language levels. These points should also be taken into account in further research. It is also worth considering the possible effects of re-testing (testing effect), when participants perform better on the post-test due to prior familiarity with the assessment structure. In addition, the influence of the Hawthorne effect cannot be completely ruled out, according to which the very fact of participation in the study and increased attention to the educational process could have a positive effect on student motivation and, accordingly, on learning outcomes. Also, the academic test was specially designed for the purposes of this study. Despite satisfactory psychometric characteristics, the use of the author's assessment tool may limit the external validity of the results due to the potential risk of constructive agreement between the test structure and the concept of the proposed AKS model.

Additional limitations are that the study was conducted only among students of humanities majors at Ukrainian higher education institutions during one academic semester. This may limit the possibility of generalizing the results to other disciplinary fields, educational systems, and longer periods of use of automated assessment systems. Further research is advisable to conduct using randomized or controlled experimental designs, using standardized assessment tools, independent samples from different universities, and testing the effectiveness of the proposed model in natural science, technical, and interdisciplinary educational contexts.

Conclusions

Thus, the study provides preliminary empirical support for the proposed three-level conceptual model of an automated knowledge assessment system (AKS) based on artificial intelligence.

The study provides preliminary empirical evidence supporting the construct- and criterion-related validity of the proposed three-level AKS model. The overall correlation with the average score of two experts was $r = 0.87$ ($p < .001$), the inter-expert agreement coefficient Cohen's $\kappa = 0.79$, F1-score = 0.89. At the same time, the somewhat lower validity for law students ($r = 0.82$) indicated the need for domain-specific fine-tuning of the model for highly specialized academic genres.

The findings suggest a significant increase in students' academic achievement: $t(249) = 31.42$, $p < .001$, $\Delta M = 13.3$ points, Cohen's $d = 0.98$. This value exceeded the values typical for studies on automated feedback and was recorded in both the subgroup of bachelors ($d = 0.94$) and masters ($d = 1.08$).

In addition, students of humanities specialties generally positively perceived the AKS model on all three dimensions of the TAM: PU = 3.91, PEOU = 3.67, OS = 3.78 (all significantly higher than the neutral value of 3.0, $p < .001$). The revealed pattern — the increase in PU and OS with an increase in the level of intellectualization of the system while simultaneously decreasing PEOU — indicated a fundamental pedagogical compromise between the semantic system and the subjective ease of interaction with it.

The study also recorded a hierarchy of the relative importance of the levels of the model in explaining the variability of educational achievements (total $R^2 = .61$): Tier 2 (dominance weight = .31; 50.8% of R^2) > Tier 3 (.19; 31.1%) > Tier 1 (.11; 18.1%).

All three levels made a statistically significant and unique contribution to the total R^2 and indicated that none of them is redundant in the structure of the model.

Thus, the scientific novelty of this study lies in the verification of a three-level hierarchical model of the AKS, which integrates the syntactic-formal, semantic-discursive, and personalized levels of intellectualization into a single architecture. Hence, the present study provides preliminary empirical support for the proposed three-tier conceptual model.

Further studies should evaluate the proposed model using randomized controlled designs, standardized assessment instruments, and diverse educational contexts to establish its generalizability and strengthen the evidence supporting the proposed architecture.

Bibliographic references

- Aquino, M. (2026). From Freire's critical pedagogy to the age of Artificial Intelligence: Continuities, ruptures, and new educational challenges. *Science Insights Education Frontiers*, 33(2), 5387–5401. <https://doi.org/10.15354/sief.26.re124>
- Basilotta-Gómez-Pablos, V., Matarranz, M., Casado-Aranda, L.-A., & Otto, A. (2022). Teachers' digital competencies in higher education: A systematic literature review. *International Journal of Educational Technology in Higher Education*, 19(1), 8. <https://doi.org/10.1186/s41239-021-00312-8>
- Byrka, M., Sushchenko, A., Luchko, V., Perun, G., & Luchko, V. (2024). Algorithmic thinking in higher education: Determining observable and measurable content. *Information Technologies and Learning Tools*, 104(6), 1–13. <https://doi.org/10.33407/itlt.v104i6.5750>

- Carvalho, R., Lins, L. F., Rodrigues, L., Miranda, P., Oliveira, H., Cordeiro, T., Bittencourt, I. I., Isotani, S., & Mello, R. F. (2024). Exploring NLP and embedding for automatic essay scoring in the Portuguese. In A. M. Olney, I.-A. Chounta, Z. Liu, O. C. Santos, & I. I. Bittencourt (Eds.), *Artificial intelligence in education. Posters and late breaking results, workshops and tutorials, industry and innovation tracks, practitioners, doctoral consortium and blue sky* (pp. 228–233). Springer. https://doi.org/10.1007/978-3-031-64315-6_18
- Cavalcanti, A. P., Barbosa, A., Carvalho, R., Freitas, F., Tsai, Y.-S., Gašević, D., & Mello, R. F. (2021). Automatic feedback in online learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence*, 2, 100027. <https://doi.org/10.1016/j.caeai.2021.100027>
- Cechinel, C., Ochoa, X., Lemos dos Santos, H., Carvalho Nunes, J. B., Rodés, V., & Marques Queiroga, E. (2020). Mapping learning analytics initiatives in Latin America. *British Journal of Educational Technology*, 51(4), 892–914. <https://doi.org/10.1111/bjet.12941>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge.
- da Silva, L. M., Dias, L. P. S., Barbosa, J. L. V., Rigo, S. J., dos Anjos, J. C. S., Geyer, C. F. R., & Leithardt, V. R. Q. (2022). Learning analytics and collaborative groups of learners in distance education: A systematic mapping study. *Informatics in Education*, 21(1), 113–146. <https://doi.org/10.15388/infedu.2022.05>
- Deroncele-Acosta, A., Sayán-Rivera, R. M. E., Mendoza-López, A. D., & Norabuena-Figueroa, E. D. (2025). Generative artificial intelligence and transversal competencies in higher education: A systematic review. *Applied System Innovation*, 8(3), 83. <https://doi.org/10.3390/asi8030083>
- Dos Santos, R. F., Da Silva, J. d. M. G., De Souza, M. H., & Borges, P. C. R. (2026). Development of an artificial intelligence model to support teaching in dynamic educational environments. *Remunom*, 13(03), 1–25. <https://doi.org/10.66104/fkfg6q38>
- Duarte Ehlert, A., Da Rocha Junqueira, J., Astrogildo de Freitas, L., & Brisolará Corrêa, U. (2025). A Review of recent advances on automatic question generation for the portuguese language. *The International FLAIRS Conference Proceedings*, 38. <https://doi.org/10.32473/flairs.38.1.138961>
- Dung, D. A., Toan, N. T., Minh, N. D., & Anh, N. Q. (2026). Reimagining quality: Artificial intelligence, governance and the politics of data in higher education. *Higher Education for the Future*, 13(1), 50–73. <https://doi.org/10.1177/23476311251386123>
- Durso, S. d. O., & Arruda, E. P. (2022). Artificial intelligence in distance education: A systematic literature review of Brazilian studies. *Problems of Education in the 21st Century*, 80(5), 679–692. <https://doi.org/10.33225/pec/22.80.679>
- Fajardo-Ramos, D. C., Chiappe, A., & Mella-Norambuena, J. (2025). Human-in-the-loop assessment with AI: Implications for teacher education in Ibero-American universities. *Frontiers in Education*, 10, 1710992. <https://doi.org/10.3389/feduc.2025.1710992>
- Falcão, T. P., Mello, R. F., Rodrigues, R. L., Diniz, J. R. B., Tsai, Y.-S., & Gašević, D. (2020). Perceptions and expectations about learning analytics from a brazilian higher education institution. In *LAK '20: 10th International Conference on Learning Analytics and Knowledge*. ACM. <https://doi.org/10.1145/3375462.3375478>
- Feitosa, R. G. F., Campos, G. A. L. d., Santos, I. d. S., Gonçalves, C. H. R., Serra, A. d. B., Oliveira, A. R. d., Feitosa, P. L. P., Santos, Y. D., Bispo, E. L., & Esmeraldo, G. Á. R. M. (2025). Competency and skill-based

- educational recommendation system. *International Journal of Artificial Intelligence in Education*, 35(1), 135-154. <https://doi.org/10.1007/s40593-024-00423-z>
- Fernandes, R. M., Nagata, V. d. M. N., Leal Filho, W., Rampasso, I. S., Avila, L. V., da Silva, J. M. N., Anholon, R., & Martins, V. W. B. (2026). Proposal of a framework to enhance the adoption of artificial intelligence in higher education considering sustainable development: a hybrid Lawshe-TOPSIS approach. *International Journal of Educational Management*, 40(1-2), 8–28. <https://doi.org/10.1108/ijem-08-2024-0441>
- Ferreira, H. N. M., Brant-Ribeiro, T., Araujo, R. D., Dorca, F. A., & Cattelan, R. G. (2017). An automatic and dynamic knowledge assessment module for adaptive educational systems. In *2017 IEEE 17th International Conference on Advanced Learning Technologies (ICALT)*. IEEE. <https://doi.org/10.1109/icalt.2017.86>
- Garcia, S., Marques, E. C. M., Mello, R. F., Gašević, D., & Falcão, T. P. (2021). Aligning expectations about the adoption of learning analytics in a brazilian higher education institution. *Lecture Notes in Computer Science* (c. 173–177). Springer International Publishing. https://doi.org/10.1007/978-3-030-78270-2_31
- Guardia-Paniura, C. H., Cueva-Luza, T., Cruz-Carpio, F. M., Ito-Díaz, R. R., Apaza-Paco, D. V., Rosas-Rojas, N., Mamani-Mamani, B., Terrero-Pérez, Á., Yaedú, R. Y. F., & Peralta-Mamani, M. (2026). Human and AI-generated feedback in higher education: A systematic review of effectiveness and student perceptions. *Contemporary Educational Technology*, 18(1), ep623. <https://doi.org/10.30935/cedtech/17863>
- Guerrero-Quíñonez, A. J., Bedoya-Flores, M. C., Mosquera-Quíñonez, E. F., Mesías-Simisterra, Á. E., & Bautista-Sánchez, J. V. (2023). Artificial Intelligence and its scope in Latin American higher education. *Ibero-American Journal of Education & Society Research*, 3(1), 264–271. <https://doi.org/10.56183/iberoeds.v3i1.627>
- Hilliger, I., Ortiz-Rojas, M., Pesántez-Cabrera, P., Scheihing, E., Tsai, Y., Muñoz-Merino, P. J., Broos, T., Whitelock-Wainwright, A., Gašević, D., & Pérez-Sanagustín, M. (2020). Towards learning analytics adoption: A mixed methods study of data-related practices and policies in Latin American universities. *British Journal of Educational Technology*, 51(4), 915–937. <https://doi.org/10.1111/bjet.12933>
- Hu, L.-t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1–55. <https://doi.org/10.1080/10705519909540118>
- Kosasi, S., Vedyanto, Kasma, U., & Ayu Eka Yuliani, I. D. (2020). The mediating role of learning analytics to improve student academic performance. *2020 2nd International Conference on Cybernetics and Intelligent System (ICORIS)*. IEEE. <https://doi.org/10.1109/icoris50180.2020.9320802>
- Leal Filho, W., Ribeiro, P. C. C., Mazutti, J., Lange Salvia, A., Bonato Marcolin, C., Lima Silva Borsatto, J. M., Sharifi, A., Sierra, J., Luetz, J., Pretorius, R., & Viera Trevisan, L. (2024). Using artificial intelligence to implement the UN sustainable development goals at higher education institutions. *International Journal of Sustainable Development & World Ecology*, 31(6), 726–745. <https://doi.org/10.1080/13504509.2024.2327584>
- Lima, Y. O. d., Silveira Braga, C. A., & Pereira, I. F. (2026). Estimating the impact of industry 4.0 automation on curricular competence indicators in brazilian vocational education and training: A mixed-methods ai-supported analysis. *International Journal for Research in Vocational Education and Training*, 13(2), 211–236. <https://doi.org/10.13152/ijrvet.13.2.3>

- Lin, F., Morland, R., & Yan, H. (2024). QuizMaster: An adaptive formative assessment system. In *Generative Intelligence and Intelligent Tutoring Systems* (p. 55–67). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-63028-6_5
- Lourenço, A. A., & Valente, S. (2026). Artificial Intelligence in learning: An integrative framework for education 4.0. *Creative Education*, 17(3), 393–422. <https://doi.org/10.4236/ce.2026.173025>
- Martinho, V. R. D. C., Nunes, C., & Minussi, C. R. (2013). An intelligent system for prediction of school dropout risk group in higher education classroom based on artificial neural networks. In *2013 IEEE 25th International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE. <https://doi.org/10.1109/ictai.2013.33>
- Mello, R. F., Neto, R., Fiorentino, G., Alves, G., Arêdes, V., Silva, J. V. G. F., Falcão, T. P., & Gašević, D. (2022). Enhancing instructors' capability to assess open-response using natural language processing and learning analytics. In *Lecture Notes in Computer Science* (pp. 102–115). Springer International Publishing. https://doi.org/10.1007/978-3-031-16290-9_8
- Oksanen, A., Osma, T., Heiskari, M., Cvetkovic, A., Ruokosuo, E. S., Koike, M., Arriaga, P., & Savolainen, I. (2026). Mapping AI learning readiness self-efficacy worldwide: Scale validation and cross-continental patterns. *Computers in Human Behavior: Artificial Humans*, 7, 100251. <https://doi.org/10.1016/j.chbah.2026.100251>
- Ramírez-Montoya, M. S., Castillo-Martínez, I. M., Sanabria-Z, J., & Miranda, J. (2022). Complex thinking in the framework of education 4.0 and open innovation—a systematic literature review. *Journal of Open Innovation: Technology, Market, and Complexity*, 8(1), 4. <https://doi.org/10.3390/joitmc8010004>
- Rodrigues, L., Toda, A. M., Palomino, P. T., Oliveira, W., & Isotani, S. (2020). Personalized gamification: A literature review of outcomes, experiments, and approaches. In *TEEM'20: Eighth International Conference on Technological Ecosystems for Enhancing Multiculturality*. ACM. <https://doi.org/10.1145/3434780.3436665>
- Ruiz-Casado, M., Alfonseca, E., & Castells, P. (2005). Automatic extraction of semantic relationships for Wordnet by means of pattern learning from Wikipedia. In *Natural Language Processing and Information Systems* (pp. 67–79). Springer Berlin Heidelberg. https://doi.org/10.1007/11428817_7
- Salas-Pilco, S. Z., & Yang, Y. (2020). Learning analytics initiatives in Latin America: Implications for educational researchers, practitioners and decision makers. *British Journal of Educational Technology*, 51(4), 875–891. <https://doi.org/10.1111/bjet.12952>
- Skrypnyk, A., Lytvyn, N., Kholod, I., Didenko, N., & Ivashchuk, A. (2024). Linguistic imagology as a new approach to the analysis of linguistic images: Methods and theoretical aspects. *Amazonia Investiga*, 13(73), 220–230. <https://doi.org/10.34069/ai/2024.73.01.18>
- Toda, A. M., Valle, P. H. D., & Isotani, S. (2018). The dark side of gamification: An overview of negative effects of gamification in education. In *Communications in Computer and Information Science* (pp. 143–156). Springer International Publishing. https://doi.org/10.1007/978-3-319-97934-2_9
- UNESCO. (2023). *New UNESCO global survey reveals impact of COVID-19 on higher education*. <https://www.unesco.org/en/articles/new-unesco-global-survey-reveals-impact-covid-19-higher-education>