International Journal of Evaluation and Research in Education (IJERE)
Vol. 15, No. 4, August 2026, pp. 2874~2882
ISSN: 2252-8822, DOI: 10.11591/ijere.v15i4.38496

Evaluating the methodological admissibility of generative AI tools for linear regression in graduate-level research

Valery Okulich-Kazarin[1,2], Kanat Kozhakhmet[1]
[1]Institute of Digital Transformation and Artificial Intelligence, Narxoz University, Almaty, Kazakhstan
[2]Institute of Management and Quality Sciences, Akademia Humanitas w Sosnowcu, Sosnowiec, Poland

Article Info

Article history:

Received Jan 2, 2026
Revised Apr 26, 2026
Accepted Jun 26, 2026

Keywords:

AI-assisted research methodology
Generative AI
Graduate education research
Linear regression analysis
Statistical reproducibility
Statistical tolerance thresholds

ABSTRACT

With the growing use of generative artificial intelligence (AI) in academia, a key methodological question concerns the statistical correctness of AI-assisted quantitative analysis. This study empirically evaluates the use of generative AI tools for linear regression in graduate-level research. The authors used a methodological approach in which estimates from four AI systems (ChatGPT 4.0, DeepSeek v3.2, Gemini 3 Pro, and Grok 4.1) were compared with estimates obtained using Microsoft Excel (Windows 10). The analysis was performed on five time series using a fixed prompt structure. Comparability was assessed using thresholds for regression coefficients, the coefficient of determination (R^2), and predicted results for 2030. The results show that under controlled conditions and within the ordinary least squares (OLS) method, the AI tools generate statistical results with varying degrees of accuracy. However, deviations in coefficients and predictions highlight the need for systematic validation. The study concludes that AI tools can serve as auxiliary methodological support, provided transparency, reproducibility, and threshold-based verification are ensured in graduate research practice.

Corresponding Author:

Valery Okulich-Kazarin
Institute of Digital Transformation and Artificial Intelligence, Narxoz University
Almaty, Kazakhstan
Email: okwalery@gmail.com

1. INTRODUCTION

Despite growing use of generative artificial intelligence (AI) in graduate research [1], [2], there is a lack of empirically testable criteria for evaluating the methodological admissibility of AI-generated statistical results. So, a methodologically significant question arises: do such tools ensure statistical correctness when performing standard quantitative analysis procedures? Several studies [3]–[5] demonstrate the sustained interest of researchers in graduate-level research and student projects. However, the problem of its methodological admissibility within the framework of classical statistical procedures remains insufficiently empirically studied.

Linear regression, widely used in graduate-level research in the fields of education, economics, engineering and social sciences, represents a more complex use case of AI tools. However, previous works [1], [2] dealt with the processing of scaled data and did not touch upon the construction of regression models. Most existing publications focus on the pedagogical aspects of AI application [6]–[16] or on the automation of educational tasks [17], [18]. At the same time, the issue of quantitative comparability of results and reproducibility of the analytical procedure when using AI remains insufficiently verified empirically [19]. Incorrect regression coefficients can affect evaluation reports, thesis assessments, policy conclusions, and student research quality. There are practically no studies based on pre-specified quantitative thresholds of

acceptable statistical deviation. The present study aims to fill this gap by employing a testable approach based on predefined quantitative thresholds.

The purpose is to assess the conditions and limits of the permissibility of using AI tools in performing standard linear regression in the context of graduate-level research, rather than to develop a new statistical procedure. AI is conceptualized strictly as a computational mechanism rather than an inferential agent. The contribution of this study lies in introducing an empirically testable framework for evaluating the methodological admissibility of generative AI tools in performing linear regression within graduate-level research. In accordance with the purpose, three research questions (RQ) were formulated.

- Do the linear regression coefficients (A, B), the determination coefficient (R^2), and the predicted values obtained using AI tools remain within the pre-established threshold of acceptable statistical deviation relative to the Microsoft Excel algorithmic benchmark calculation (Trendline OLS)? (RQ1)

The following acceptable threshold values are adopted: deviation of the regression coefficient less than ±1% from the reference value; deviation of R^2 less than ±0.01; deviation of the predicted value less than ±5%. Exceeding the threshold values is considered a violation of the principle of methodological equivalence.

- Is the structural reproducibility of the ordinary least squares (OLS) procedure maintained when using different AI tools with a fixed prompt and identical input dataset? (RQ2)

The reproducibility of a computational procedure is defined as: correct description of the linear regression $Y = A \times X + B$; calculation and output of the coefficients A and B; calculation and output of R^2; calculation and output of the predicted result for 2030. Violation of each element is considered a violation of reproducibility.

- Under what methodological conditions can the use of AI be considered acceptable without violating the requirements of statistical correctness and academic responsibility? (RQ3)

Acceptability is considered to be achieved when the following conditions are simultaneously met: compliance of the results with the threshold values; transparency of the structure of the request; check of the results by the researcher; maintenance of the researcher's responsibility for interpretation.

The methodology bases on the AI-assisted research methodology approach [19]–[22]. Within this approach, AI tools are viewed strictly as mechanisms that support the execution of established analytical procedures, without changing their methodological or epistemological foundations. AI prompts [1], [2] are conceptualized as additional computational tools for calculating standard statistical coefficients [23], [24]. Their use is accepted only to the extent that the classical requirements of linear regression analysis are preserved.

The research framework is based on three interrelated principles: i) the principle of methodological equivalence (results obtained using AI should remain quantitatively comparable with reference calculations); ii) the principle of procedural reproducibility (the use of AI tools should preserve the structure and logical consistency of the OLS procedure); and iii) the principle of limited applicability (AI tools can assist in calculations, but do not replace the researcher and their interpretive responsibility). This concept does not aim to revise educational methodology [25], [26] or statistical theory [23], [24]. Its purpose is to define the admissibility and limitations of using AI in research [27] under clearly defined methodological conditions. First of all, at the graduate-level [19], [22].

2. METHOD

2.1. Research design and analytical procedure

The analysis focused on verifying whether the AI tools reproduce the standard algorithmic implementation of the OLS method and whether the obtained results remain within the acceptable statistical deviation relative to the benchmark calculation. The built-in linear trendline function of Microsoft Excel (Insert → Chart → Add Trendline → Display Equation and R^2), executed in Windows 10, was used as the benchmark algorithmic calculation. This tool implements the standard OLS algorithm for estimating the coefficients of a linear model and calculating the coefficient of determination (R^2).

The empirical basis consisted of five time series [1], differing in the length of observations (from 7 to 21 units) and the dynamics of the indicators. These series are taken from the author's research practice, the first of which was published in the paper [7]. The range of values of the dependent variables varied in scale between the series. This made it possible to test the robustness of the reproducibility of OLS at different orders of magnitude. Extreme outliers and structural breaks were not detected. For each data set, an identical structure of the analytical procedure and a fixed query formulation were used when accessing AI tools. Moreover, the basic assumptions of classical single-factor linear regression (linearity of the relationship, independence of observations, and stability of the error variance) were observed. In-depth analysis of

residuals and testing for heteroscedasticity were not performed. At the same time, parametric robustness was tested through a comparison of coefficients and predicted values within the statistical tolerance framework.

The analytical process for processing each data series included: formalization of a linear model of the form $Y = A \times X + B$; estimation of coefficients A and B; calculation of the coefficient of determination (R^2), t-statistic, and p-value. The comparability of the results was assessed within a pre-defined system of statistical tolerances (statistical tolerance framework). Thus, the study was designed as a testable model: if the results fell outside the established limits, the assumption of statistical comparability would be considered refuted. This ensures the falsifiability of the methodological statement and reduces the risk of uncritical interpretation of numerical similarity as evidence of correctness [27], [28]. Responsibility for the interpretation of the results and control over the correctness of the calculations remained with the researcher at all stages of the analysis.

2.2. AI tools and comparison strategy

The authors used four generative models: ChatGPT 4.0, DeepSeek v3.2 (DeepSeekMath-Base version 7B), Gemini 3 Pro, and Grok 4.1. All tools were publicly available, allowing students to apply AI tools in their graduate-level research. To minimize variability, a fixed query structure was used: use an identical source dataset; instruct the model to perform OLS linear regression for $Y = A \times X + B$; output A, B, R^2, t-statistic, and p-value; and instruct the model not to apply additional data transformations. The prompt did not contain any instructions on the interpretation of the results and did not include any evaluative statements. The following parameters were used for comparison: A and B coefficients; the coefficient of determination (R^2); the presence and accuracy of the level of statistical significance (p-value); and the predicted value for 2030 (manually calculated for the selected five-year horizon to 2030).

The built-in trendline function in Microsoft Excel calculates A, B, and R^2 coefficients, but does not calculate p-values. Therefore, p-values were used solely as an additional indicator of the procedural completeness of the AI tools. They were not used as an element of direct quantitative comparison with the benchmark. Comparisons were conducted within the statistical deviation thresholds described before. The authors did not aim to identify a "best" AI tool and did not interpret minimal numerical differences as a basis for ranking models. The comparison strategy focused on verifying the conditions of methodological equivalence and procedural reproducibility when using the AI tools.

2.3. Statistical tolerance framework

To ensure the verifiability of the results and prevent arbitrary interpretation of numerical similarities, a system of statistical tolerances was introduced in the study. Acceptable deviation limits are set as [23], [24]: for the regression coefficients (A, B) – less than ±1% from the reference value; for the determination coefficient (R^2) – less than ±0.01; for the predicted value – less than ±5%. The system of statistical tolerances serves as a criterion for assessing the methodological equivalence of the results obtained using AI tools relative to the reference calculation (Windows 10).

The threshold values were selected to balance computational sensitivity and interpretative stability of the model. Minor discrepancies in the results may arise due to rounding and calculation precision. However, they should not alter the trend slope, the coefficient of determination, or the forecast result. Deviations exceeding the threshold values are considered a violation of the principle of methodological equivalence. Exceeding the threshold values constitutes grounds for rejecting the assumption of statistical comparability. Thus, the threshold system serves to verify the assumption that methodological equivalence is considered confirmed only if the quantitatively specified threshold values are met.

2.4. Reproducibility, hallucination risk and methodological control

Reproducibility of the procedure is defined as the simultaneous fulfillment of the following conditions: correctness of the model $Y = A \times X + B$; no changes to the input data; calculation of the coefficients A and B using standard OLS logic; calculation of the coefficient of determination (R^2); and correctness of the forecast for 2030. Numerical discrepancies within the limits set within the threshold values are not interpreted as a violation of reproducibility. This is crucial to consider the risk of stochastic variability in generative models. The architecture of large language models is probabilistic in nature. This means that even with the same query, minor variations in the numerical results are possible.

The study used a fixed query structure, and calculations were performed with minimal variability in the generation parameters. If significant deviations were identified, the results were re-checked. In the context of this study, AI hallucination is defined as the generation of statistically plausible but computationally incorrect results due to the probabilistic nature of the language model. To minimize this risk, the following methodological control measures were applied: use of a reference algorithmic calculation for comparison; introduction of thresholds for acceptable deviations; fixed query structure; avoidance of

interpretive instructions in the task formulation; mandatory verification of results by the researcher. This study does not assume epistemic equivalence between generative models and specialized statistical software. AI tools are considered purely as probabilistic systems requiring external algorithmic verification.

3. RESULTS AND DISCUSSION

3.1. Results related to RQ1

The results of calculating the linear regression coefficients to answer RQ1 are presented in Table 1. At first glance, the data in Table 1 demonstrate the numerical comparability of the results obtained using various AI tools with the benchmark algorithmic calculation in Microsoft Excel (Trendline OLS). However, a proper assessment of comparability requires a quantitative analysis of deviations relative to established statistical deviation thresholds. The corresponding calculations are presented in Table 2.

Table 1. Linear regression analysis for example-1 using various AI tools

Indicator	ChatGPT 4.0	DeepSeek v3.2	Gemini 3 pro	Grok 4.1	Microsoft Excel
Regression coefficient A	10308.00	10308.00	10308.00	9541.80	10308.00
Regression coefficient B	107069.00	96761.00	96760.90	104722.70	96761.00
Determination coefficient R^2	0.9768	0.9765	0.9889	0.9734	0.9768
Significance level p	0.0000	0.0000	0.0000	0.0002	-
Forecast for 2030	251381.00	241073.00	241073.20	228796.40	241073.00

Table 2 shows that DeepSeek v3.2 fully complies with the established thresholds. All indicators are within the acceptable statistical deviation, confirming methodological equivalence for this example. ChatGPT 4.0 demonstrates an excess of the acceptable deviation for the absolute term (B) by more than 10%. However, the predicted value remains within ±5%, indicating a partial violation of parametric stability while maintaining practical comparability of the forecast. Gemini 3 Pro exceeds the acceptable threshold for R^2 (Δ=0.0121>0.01). Despite maintaining high explanatory power of the model, the formal equivalence criterion is not met. Grok 4.1 demonstrates deviations in several parameters, including coefficients and predicted value, which does not allow us to confirm methodological equivalence for this example. Thus, the results of RQ1 indicate a differentiated nature of statistical comparability. While maintaining the structural correctness of the model, the quantitative stability of the parameters depends on the AI tool used.

Table 2. Deviation analysis relative to benchmark calculation (Example-1)

Tool	ΔA (%)	ΔB (%)	ΔR^2	ΔForecast (%)	Within tolerance?
ChatGPT 4.0	0.00	+10.64	0.0000	+4.27	Partial (B)
DeepSeek v3.2	0.00	0.00	0.0000	0.00	Yes
Gemini 3 Pro	0.00	−0.00	+0.0121	+0.00	No (R^2)
Grok 4.1	−7.44	+8.22	−0.0034	−5.09	No
Excel (Benchmark)	–	–	–	–	–

3.2. Results related to RQ2

To find the answer to RQ2, the following four examples were used, as seen in Table 3. The data presented in Table 3 demonstrate statistical coefficients and predicted values for linear regression using AI prompts across four additional empirical examples with different data characteristics. Reproducibility analysis was performed in two stages: structural reproducibility testing (full OLS procedure); quantitative assessment of deviations relative to the Microsoft Excel reference calculation within the statistical tolerance framework.

In all examples, the neural networks correctly formalized the $Y = A \times X + B$ model and showed coefficients and R^2, as shown in Table 4. In terms of the OLS procedure structure, reproducibility is confirmed. The picture regarding quantitative stability is significantly more complex. For example, for DeepSeek v3.2, deviations remain within the established thresholds in all four examples (Table 4). Thus, methodological equivalence is consistently confirmed.

For ChatGPT 4.0 and Gemini 3 Pro, significant deviations in the intercept (B) are observed in most examples (Table 4). However, the slope coefficients (B) are virtually identical. R^2 remains stable in most cases. Predicted values are often within ±5% or close to the limit. This indicates partial parametric instability while maintaining the overall structure of the linear relationship.

For Grok 4.1, example-3 demonstrates a fundamentally different pattern of deviations: $\Delta A = +13.12\%$, $\Delta B = +421.34\%$, $\Delta R^2 = +0.1958$, ΔForecast = +29.89% (Table 4). This may be due to

structural deviation of the model. In this case, R^2 increases from 0.7655 to 0.9513. This significantly changes the interpretation of the explanatory power of the model. In this case, the statistical tolerance framework records a violation of methodological equivalence. Thus, the reproducibility of the procedure does not mean the quantitative stability of the parameters.

The difference between structural reproducibility (the formal implementation of OLS) and parametric stability (compliance with established tolerances) was demonstrated in Table 4. Structural reproducibility was confirmed for all AI tools. Quantitative stability is specific. DeepSeek demonstrates stability. ChatGPT and Gemini demonstrate partial stability. Grok 4.1 demonstrates instability with certain datasets. Thus, RQ2 receives a differentiated answer: AI tools reproduce the analytical structure of OLS, but numerical stability varies and requires mandatory external verification.

Table 3. Testing the reproducibility of analytical procedures for four various AI tools (examples 2–5)

Tool	A	B	R^2	p-value	Forecast for 2030
Example-2					
ChatGPT 4.0	757.90	-164.90	0.8842	0.0000	28635.30
DeepSeek v3.2	757.80	-921.30	0.8842	0.0000	27875.10
Gemini 3 pro	757.90	-164.90	0.9043	0.0000	28635.30
Grok 4.1	757.90	-164.90	0.8842	0.0000	28635.30
Excel (Benchmark)	757.90	-922.80	0.8842	-	27877.40
Example-3					
ChatGPT 4.0	5862.20	2407.90	0.7655	0.0000	107927.50
DeepSeek v3.2	5861.60	-3450.50	0.7654	0.0000	102076.00
Gemini 3 pro	5861.60	2411.20	0.7654	0.0000	107920.00
Grok 4.1	6748.50	11099.20	0.9513	0.0000	132572.20
Excel (Benchmark)	5862.20	-3454.30	0.7655	-	102065.30
Example-4					
ChatGPT 4.0	1168.80	224.20	0.8901	0.0001	17756.20
DeepSeek v3.2	1168.80	-944.60	0.8898	0.0001	16587.40
Gemini 3 pro	1168.80	-944.60	0.8844	0.0007	16587.40
Grok 4.1	1168.80	224.20	0.8901	0.0001	17756.20
Excel (Benchmark)	1168.80	-944.60	0.8901	-	16587.40
Example-5					
ChatGPT 4.0	327.80	4025.30	0.9333	0.0000	8942.30
DeepSeek v3.2	327.80	3697.50	0.9330	0.0000	8614.50
Gemini 3 pro	327.80	4025.30	0.9333	0.0000	8942.30
Grok 4.1	354.60	4216.00	0.9440	0.0000	9535.00
Excel (Benchmark)	327.80	3697.50	0.9333	-	8614.50

Table 4. Summary of deviations relative to benchmark (examples 2–5)

Tool	ΔA (%)	ΔB (%)	ΔR^2	ΔForecast (%)	Within tolerance?
Example-2					
ChatGPT 4.0	0.00	+82.12	0.0000	+2.72	No (B)
DeepSeek v3.2	0.00	0.16	0.0000	−0.01	Yes
Gemini 3 Pro	0.00	+82.12	+0.0201	+2.72	No (R^2, B)
Grok 4.1	0.00	+82.12	0.0000	+2.72	No (B)
Example-3					
ChatGPT 4.0	0.00	+169.70	0.0000	+5.75	No
DeepSeek v3.2	−0.01	0.11	0.0001	+0.01	Yes
Gemini 3 Pro	−0.01	+169.84	0.0001	+5.74	No
Grok 4.1	+13.12	+421.34	+0.1958	+29.89	No
Example-4					
ChatGPT 4.0	0.00	+123.73	0.0000	+7.05	No
DeepSeek v3.2	0.00	0.00	0.0003	0.00	Yes
Gemini 3 Pro	0.00	0.00	−0.0064	0.00	Yes
Grok 4.1	0.00	+123.73	0.0000	+7.05	No
Example-5					
ChatGPT 4.0	0.00	+8.87	0.0000	+3.80	No (B)
DeepSeek v3.2	0.00	0.00	0.0003	0.00	Yes
Gemini 3 Pro	0.00	+8.87	0.0000	+3.80	No (B)
Grok 4.1	+8.18	+14.02	+0.0107	+10.68	No

3.3. Results related to RQ3

The empirical results allow us to infer the conditions for acceptable use of AI [19] based on the observed deviations (Tables 2 and 4). The empirical results allow us to conclude that the analysis demonstrated three distinct empirical patterns: complete parametric stability within the given tolerance

thresholds (DeepSeek v3.2 in all examples); partial parametric instability primarily affecting the intercept values (ChatGPT 4.0 and Gemini 3 Pro); and structural deviation beyond the tolerance thresholds simultaneously affecting the slope, R^2, and the predicted outcome (Grok 4.1 in example 3). These patterns allow us to formulate four empirically based tolerance conditions.

- Condition 1: quantitative validation. The presence of threshold violations in a number of AI tools indicates that the regression parameters generated by the AI cannot be accepted without verification by the researcher. The use of AI tools is permissible only if: i) the coefficients and predicted values are close to the reference calculations; ii) the deviations are estimated within predetermined tolerances.
- Condition 2: sensitivity to coefficient B. In several examples (Tables 2 and 4), deviations most often affected the intercept (B), while the slope (A) often remained stable. A change in the intercept can significantly affect forecast values. Thus, checking the stability of the intercept and independent confirmation of forecast calculations are required when deviations exceed ±1%.
- Condition 3: R^2 stability threshold. In example 3, the Grok 4.1 neural network demonstrates that a significant deviation in R^2 (+0.1958) can fundamentally change the interpretation of the model. In such cases, the use of AI cannot be considered methodologically equivalent. Therefore, the acceptability of use requires deviations in R^2 within ±0.01.
- Condition 4: researcher responsibility and control. Empirical data confirm that AI tools reproduce the computational structure, but may vary in numerical precision. Therefore, the acceptability of use is conditioned by: i) the researcher's ability to interpret statistical parameters [22]; ii) verification of results; iii) a refusal to regard AI results as epistemically autonomous. Taken together, the obtained results demonstrate that AI prompts can be considered an acceptable means of methodological support, but not as a standalone analytical tool [20].

3.4. Discussion

The obtained results allow us to clarify the methodological status of using AI in graduate-level research [1], [2] for linear regression analysis. First, the results of RQ1 and RQ2 demonstrate that generative AI tools reproduce the structure of the OLS procedure. However, their numerical stability is not constant. Full parametric equivalence was not confirmed for all models. Therefore, methodological equivalence is conditional and depends on compliance with established statistical deviation thresholds. Second, the identified deviations in the intercept and, in some cases, in R^2 demonstrate that structural reproducibility does not guarantee parametric stability. This distinction is of fundamental importance for graduate-level research, where coefficient values directly influence the study's conclusions. Third, the results of RQ3 confirm the principle of limited admissibility: AI tools can be used as an auxiliary computational mechanism, provided that they are quantitatively verified, the procedure is transparent, and researcher control is maintained [6], [22]. Empirical data do not support their epistemic autonomy.

The identified distinction between structural reproducibility and parametric stability clarifies the methodological status of AI in educational research. A number of publications consider neural networks as a tool for optimization of academic activity [9], [11], [14], [29]. However, the obtained empirical data demonstrate that computational support is not identical to the stability of statistical results. As for educational evaluation, even minor deviations in AI-generated regression outputs may influence the assessment of graduate research quality and the validity of thesis conclusions. This point becomes particularly relevant for empirical educational studies where statistical results directly inform pedagogical recommendations. For example, a comparative study of media students in different countries statistically demonstrated a significant preference for visual learning methods over auditory ones, which led to concrete recommendations for adapting lecture formats in higher education [30]. Thus, the study shifts the discussion about the use of AI in education from the level of technological efficiency [13]–[18], [25], [26] to the level of methodological admissibility. AI tools can be integrated into the research infrastructure of graduate-level education [19]–[22], but their use requires maintaining the responsibility of the researcher.

An additional limitation is the opacity of the generative model architecture (model opacity) [31]. The user has no access to the internal algorithmic processing mechanisms and possible hidden computational stages. Given this opacity, external quantitative verification of the results becomes a necessary condition for methodological reliability. The approach is validated when parametric stability conditions are met [27], but does not automatically imply the methodological equivalence of all AI tools. Empirical data clarify the applicability of AI-assisted research methodology in the context of quantitative analysis.

4. CONCLUSION

The results shift the discussion of AI use in graduate-level research from the level of technological effectiveness to the level of methodological admissibility. The empirical data clarify the limits of the

application of AI-assisted research methodology in quantitative analysis: its use is permissible only if the conditions of parametric stability and research control are met. The automatic methodological equivalence of generative models is not confirmed.

In terms of practical significance, the study's results allow for the formulation of cautious methodological recommendations how to use AI tools when performing linear regression in graduate-level research. First, the use of AI is only permissible if the obtained coefficients and predicted values are quantitatively verified within predetermined statistical deviation tolerances. Second, the coefficient (B) requires special attention, as this parameter exhibits the greatest variability and may influence predictive results. Third, if the determination coefficient (R^2) exhibits significant deviations, using the AI results without further verification is methodologically unacceptable. This may affect the interpretation of the model's explanatory power. The obtained results are contextual in nature. They relate exclusively to single-factor linear regression with a fixed prompt structure. They do not imply a ranking of AI tools outside this methodological context.

There are several limitations of the study. First, the analysis was limited to a single-variate linear regression study. The results obtained cannot be automatically generalized to multivariate or nonlinear models. Such models have higher parametric sensitivity and computational complexity. Second, the study focused on the methodological validity of the calculations. The study was not on the selection of the best AI tool or the pedagogical effects of AI use. The impact of AI application on the development of master's students' research competencies requires separate analysis. Third, the results were obtained with a fixed query structure and specific versions of the AI tools. Given the dynamic nature of neural network development, subsequent versions may exhibit a different degree of parametric stability. However, this possibility requires empirical verification and cannot be assumed normatively. A promising direction for further research is a longitudinal analysis of the stability of results when updating AI tools, as well as an assessment of their applicability in more complex statistical procedures.

FUNDING INFORMATION

This research has been funded by the grant "Application of hybrid swarm intelligence algorithms in the development of proactive multi-agent systems for the digital educational environment" of the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (No. AP26196023).

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Valery Okulich-Kazarin	✓	✓		✓	✓	✓			✓	✓	✓	✓	✓	
Kanat Kozhakhmet			✓	✓	✓		✓	✓	✓	✓			✓	✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The authors confirm that the data supporting the findings of this study are available within the article and its supplementary materials.

REFERENCES

[1] V. Okulich-Kazarin, "New ChatGPT 3.5 Instruction (Prompt) to Calculate Statistical Indicators for Student Graduation Projects," *WSEAS Transactions on Computer Research*, vol. 12, pp. 307–317, 2024, doi: 10.37394/232018.2024.12.30.

[2] V. Okulich-Kazarin, "Statistics Using Neural Networks in the Context of Sustainable Development Goal 9.5," *Sustainability*, vol. 16, no. 19, 2024, doi: 10.3390/su16198395.

[3] H. H. Yang and Y.-T. Lin, "How Knowledge Sharing and Cohesion Become Keys to a Successful Graduation Project for Students from Design College," *SAGE Open*, vol. 12, no. 3, 2022, doi: 10.1177/21582440221121785.

[4] J. O. Conner, "Student Engagement in an Independent Research Project: The Influence of Cohort Culture," *Journal of Advanced Academics*, vol. 21, no. 1, pp. 8–38, 2009, doi: 10.1177/1932202X0902100102.

[5] R. Herzallah, "Practically Oriented Design Projects in Mechatronics Engineering: A Case Study of Design Experiences and Outcomes," *International Journal of Electrical Engineering & Education*, vol. 47, no. 1, pp. 31–46, 2010, doi: 10.7227/IJEEE.47.1.4.

[6] M. P. Babanoğlu, T. Ö. Karataş, and E. Dündar, "Envisioning the Future of AI-Assisted EFL Teaching and Learning: Conceptual Representations of Prospective Teachers," *SAGE Open*, vol. 15, no. 2, 2025, doi: 10.1177/21582440251341590.

[7] A. Alkan, "Analysis of Factors Affecting Academic Success with Machine Learning: Data-Driven Inferences in Education," *Turkish Journal of Engineering*, vol. 10, no. 1, pp. 48–62, Dec. 2025, doi: 10.31127/tuje.1779491.

[8] S. Namboothiri, T. Varghese, M. Jacob, S. Job, and J. Cyriac, "Integrating Artificial Intelligence with NHEQF Descriptors for Pedagogical Excellence," *Higher Education for the Future*, vol. 12, no. 1, pp. 27–50, 2025, doi: 10.1177/23476311241290894.

[9] D. Saccenti, M. Buattini, S. Grazioli, and D. Torres, "Navigating the AI frontier: Should we fear ChatGPT use in higher education and scientific research? Finding a middle ground through guiding principles and practical applications," *Possibility Studies & Society*, vol. 2, no. 4, pp. 415–437, 2024, doi: 10.1177/27538699241231862.

[10] S. O. Sampson, H. Keene, A. Langworthy, O. Omotilewa, M. Paul, and G. Shukla, "Artificial intelligence as an academic ally: A quality improvement initiative for a graduate evaluation course," *American Journal of Evaluation*, vol. 47, no. 1, pp. 8–17, 2026, doi: 10.1177/10982140251400866.

[11] T. Dogru *et al.*, "The implications of generative artificial intelligence in academic research and higher education in tourism and hospitality," *Tourism Economics*, vol. 30, no. 5, pp. 1083–1094, 2023, doi: 10.1177/13548166231204065.

[12] X. Yang, Q. Wang, and J. Lyu, "Assessing ChatGPT's Educational Capabilities and Application Potential," *ECNU Review of Education*, vol. 7, no. 3, pp. 699–713, 2024, doi: 10.1177/20965311231210006.

[13] Q. Wang, Z. Ma, G. Zhang, and Q. Miao, "Digital Transformation Driving Educational Reforms in Higher Education: A Paradigm Shift in Teaching and Learning," *Beijing International Review of Education*, vol. 7, no. 4, pp. 254–272, 2025, doi: 10.1177/25902547251395880.

[14] J.-M. Aguado-García, S. Alonso-Muñoz, and C. De-Pablos-Heredero, "Using Artificial Intelligence for Higher Education: An Overview and Future Research Avenues," *SAGE Open*, vol. 15, no. 2, 2025, doi: 10.1177/21582440251340352.

[15] H. Judson, "'But I Didn't Use ChatGPT!': Democratic Course Design and Generative Artificial Intelligence in Higher Education Landscape," *Teaching Sociology*, vol. 53, no. 3, pp. 246–256, 2025, doi: 10.1177/0092055X251342530.

[16] J.-C. Ruano-Borbalan, "The transformative impact of artificial intelligence on higher education: A critical reflection on current trends and futures directions," *International Journal of Chinese Education*, vol. 14, no. 1, 2025, doi: 10.1177/2212585X251319364.

[17] H. A. Hamzah, M. S. Abu Seman, and M. Ahmed, "The impact of artificial intelligence in enhancing online learning platform effectiveness in higher education," *Information Development*, vol. 41, no. 3, pp. 794–810, 2025, doi: 10.1177/02666669251315842.

[18] T. Ekundayo and S. A. Chaudhry, "Mapping the Future: A Bibliometric Analysis of Engagement Trends in Artificial Intelligence within Higher Education," *Digital Technologies Research and Applications*, vol. 4, no. 3, pp. 1–21, 2025, doi: 10.54963/dtra.v4i3.1285.

[19] J. Wilsdon *et al*., *The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management*, 2015, doi: 10.13140/RG.2.1.4929.1363.

[20] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. Pearson Education, 2021.

[21] S. Hajkowicz *et al.*, *Artificial Intelligence for Science: Adoption Trends and Future Development Pathways*. CSIRO Data61, Brisbane, Australia, 2022.

[22] W. Holmes, M. Bialik, and C. Fadel, *Artificial Intelligence in Education: Promise and Implications for Teaching and Learning*. Center for Curriculum Redesign, Boston, MA, USA, 2019.

[23] M. Çetinkaya-Rundel and J. Hardin, *Introduction to Modern Statistics.* OpenIntro, 2024.

[24] D. Singpurwalla, *A Handbook of Statistics: An Overview of Statistical Methods*. Bookboon, London, UK, 2015.

[25] B. K. Daniel and T. Harland, *Higher Education Research Methodology: A Step-by-Step Guide to the Research Process*, 1st ed. Routledge, London, UK, 2017, doi: 10.4324/9781315149783.

[26] E. Cleaver, M. Lintern, and M. McLinden, *Teaching and Learning in Higher Education: Disciplinary Approaches to Educational Enquiry*, 2nd ed. SAGE Publications Ltd., London, UK, 2018.

[27] O. Boursalie, R. Samavi, and T. E. Doyle, "Evaluation methodology for deep learning imputation models," *Experimental Biology and Medicine*, vol. 247, no. 22, pp. 1972–1987, 2022, doi: 10.1177/15353702221121602.

[28] M. Shen, J. Wu, C. Patt, and R. Mendoza-Denton, "Understanding graduate school through AI: A scalable approach to thematic coding," *International Journal of Qualitative Methods*, vol. 24, 2025, doi: 10.1177/16094069251395943.

[29] A. Verma and A. S. Dhaigude, "Bridging theory and practice: A mixed-method exploration of AI's role in higher education," *SAGE Open*, vol. 16, no. 1, 2026, doi: 10.1177/21582440261417610.

[30] V. Okulich-Kazarin, "What Method of Learning do Media Students Prefer at Lectures: Auditory or Visual?" *Universal Journal of Educational Research*, vol. 8, no. 6, pp. 2660–2667, 2020, doi: 10.13189/ujer.2020.080650.

[31] K. M. L. Jones, "Generative AI in qualitative research and related transparency problems: A novel heuristic for disclosing uses of AI," *International Journal of Qualitative Methods*, vol. 24, 2025, doi: 10.1177/16094069251404329.

BIOGRAPHIES OF AUTHORS

Valery Okulich-Kazarin is a Professor at the Akademia Humanitas w Sosnowcu, Sosnowec (Poland) and a Leading Researcher at the Narxoz University, Almaty (Kazakhstan). He received his Doctor of Habilitation degree in 2003. His specialization is management of higher education services. In the early 1990s, he began using artificial intelligence tools to improve the quality of teaching. Over the past 10 years, students under his supervision have won International scientific competitions more than 30 times. In 2018, he was awarded the Diploma "Best University Teacher-2018". In 2020, he received the highest academic grade "Doktor Honoris Causa". In 2025, he received a grant on the topic of artificial intelligence. He can be contacted at email: okwalery@gmail.com.

Kanat Kozhakhmet is the President of Narxoz University, has a PhD degree in Informatics, Computer Engineering and Control, teaching and research experience, in particular, supervising doctoral and master's dissertations. He has experience in leading grant and program-targeted funding projects of the Ministry of Higher Education of the Republic of Kazakhstan: AP05133600 "Developing and implementing innovative competence-based model of multilingual IT specialist in the course of national education system modernization" (2018-2020); AP26196023 "Application of hybrid swarm intelligence algorithms in the development of proactive multi-agent systems for the digital educational environment". He is the author of more than 50 scientific papers, including 3 monographs on the competence approach in IT education and 18 publications indexed in the Scopus database. He can be contacted at email: kanat.kozhakhmet@narxoz.kz.