

A.U. Serikbayeva^{1*}, D.M. Akizhanova²¹Al-Farabi Kazakh National University, Almaty, Kazakhstan²L.N. Gumilyov Eurasian National University, Astana, Kazakhstan

*e-mail: assiyaserikbayeva237@gmail.com

¹ORCID 0009-0002-7516-482X, ²ORCID 0000-0002-1300-0339

CAN GENERATIVE AI REPLACE HUMAN ESSAY RATERS? A PSYCHOMETRIC COMPARISON OF GPT-5.2, GEMINI, AND TEACHER RATINGS

Evaluating student writing is a crucial yet time-consuming process. Large Language Models (LLMs) and other recent advancements in generative AI provide potential solutions to support teachers in essay assessment. This study psychometrically evaluates large language models such as GPT-5.2 and Gemini with human raters to examine reliability of these solutions. All 55 essays were collected and assessed by GPT 5.2, Gemini and an experienced teacher using a four-criterion scoring rubric. The aim of this study is to psychometrically compare the results of human raters, GPT-5.2, and Gemini assessments on English essays, to determine their rating reliability, intercorrelation, rating differences, and the construct validity of the rating rubric. The results demonstrated that GPT-5.2 assigned the highest vocabulary scores, while Gemini produced the highest grammar scores. Overall, GPT-5.2 showed high agreement with teacher ratings and can be used as a reliable adjunct to essay grading. However, teacher professional judgment remains important when assessing complex content and contextual features. These findings support the use of generative AI as a supplementary tool for improving the efficiency and consistency of essay assessment in EFL contexts.

Keywords: automated essay scoring, Large Language Models (LLMs), GPT-5.2, Gemini, human rater, assessment; psychometric assessment.

Introduction

Writing assessment is a crucial component of language education. It is not only a tool for determining the level of linguistic competence, but also an important pedagogical process that allows to improve the quality of teaching, monitor the learning outcomes of students, and provide effective feedback [1].

Although traditional essay assessment is one of the primary methods for assessing students' writing competence, a number of issues remain regarding its reliability and fairness. First, human ratings are influenced by factors such as the assessor's experience, linguistic and cultural characteristics, and personal perceptions of the quality of written work. Such differences can lead to inconsistencies between assessors and variability in assessment outcomes [2].

Additionally, manually grading essays is time-and labour-intensive. Human evaluation rigor varies when grading essays. Such differences threaten the validity and fairness of writing assessment [3].

Advancements in Artificial Intelligence (AI) and large language models (LLMs) such as GPT-5.2 and Gemini offer a new opportunity for educational assessment.[2] These models are capable of analysing and generating natural language with high proficiency, also they can evaluate time consuming assignments and provide feedback [4].

Against this backdrop, large language models (LLMs) have been widely used in the field of automatic essay grading. Previous studies have shown that large language models such as GPT-5.2 can closely match the results of human graders, it has not yet been fully proven that they perform equally well across all grading criteria. Therefore, a comprehensive psychometric assessment of the essay assessment capabilities of large language models is one of the most pressing issues in contemporary educational research. While many studies have shown that GPT-5.2 or other large language models have a high level of agreement with human evaluation, several important gaps

remain. First, most studies have considered only one large language model and have not compared the evaluation results of multiple AI systems [4, 5, 6]. Moreover, there are very few empirical studies in the EFL (English as a Foreign Language) context in Kazakhstan comparing GPT-5.2, Gemini, and teacher assessments from a psychometric perspective [7, 8]. Therefore, the aim of this study is to psychometrically compare the results of human raters, GPT-5.2, and Gemini assessments on English essays, to determine their rating reliability, intercorrelation, rating differences, and the construct validity of the rating rubric. In this regard, the objectives of the study are

To compare the descriptive indicators of the Teacher, GPT-5.2 and Gemini ratings.

To determine the reliability and construct validity of the rating rubric.

To assess the inter-rater agreement and differences from a psychometric perspective.

Overall, these research questions aim to comprehensively evaluate the psychometric characteristics of GPT-5.2, Gemini, and teacher raters, and provide scientific evidence for the reliable and valid use of generative artificial intelligence in grading English essays.

Literature review

In recent years, the potential of large language models (LLMs) for assessing written work has been widely investigated. García-Varela et al. [9] studied the stability and fairness of the GPT-5.2 system in assessing essays and showed that the results of the AI assessment directly depend on the quality of the rubric used and the prompt structure. The results of the study showed that when assessing with simple instructions, the grades were less stable, but when using a specific rubric, assessment examples and a standardized assessment format, the assessment results were more stable and consistent. The authors proposed methods such as decision tables, rubric normalization and prompt engineering to improve the reliability of the assessment.

Moreover, Seßler et al. [10] compared the results of GPT-3.5, GPT-4, o1, LLaMA, and Mixtral models in assessing student essays in German with the ratings of 37 teachers. The study found that while the closed models (especially o1 and GPT-4) showed high agreement with the teacher ratings, the open models showed lower agreement. Furthermore, while the results of the models on linguistic criteria were closer to the teachers, differences remained on the content criteria.

Karaçeper and Kıray [11] investigated the ability of the ChatGPT-4o model to automatically grade essays in English and found that it only showed fair agreement with human raters. The authors concluded that although ChatGPT-4o was able to perform the grading quickly and objectively, it did not fully account for complex linguistic features and pedagogical considerations, and therefore could not fully replace human raters in high-stakes grading.

Wu et al. [12] found that an automated essay grading system based on GPT demonstrated high scoring consistency compared to human raters. Using the Many-Facet Rasch Measurement (MFRM) model, the authors showed that GPT grades, while having high internal consistency, differed from human graders on more complex criteria such as content. They also suggested that GPT could be used as a supplementary tool to enhance grading reliability, rather than as a complete replacement for human grading.

Within the Kazakhstani context, research on AI-based development and evaluation of Unified National Testing (UNT) items measurement was conducted by Abdrasilov et. al [8] that the most AI-generated tasks were considered suitable after expert review To promote a human-in-the-loop approach to educational evaluation, the authors stressed that expert validation was still crucial.

Yermaganbetova et al. [13] studied the effectiveness of an AI-integrated educational platform in higher education and showed that artificial intelligence improved students' academic achievement, learning motivation, and self-regulated learning skills through automated assessment, personalized feedback, and adaptive learning. In the study, GPT was used primarily for question generation and formative feedback, while Gemini was used for rubric-based assessment of practical work. The authors also noted that AI assessment should be viewed as a tool to support, rather than completely replace, the teacher's professional judgment.

A review of the literature shows that most previous studies have compared a single large language model, such as GPT, with human raters. However, there is still a lack of research that

compares multiple modern generative AI models, such as GPT-5.2 and Gemini, with teacher ratings on the basis of psychometric performance simultaneously. Furthermore, inter-rater agreement, internal reliability, and criterion-referenced differences in ratings have not been comprehensively addressed.

Materials and research methods

This research employed a quantitative research design to compare human ratings with Large Language Models (LLMs) such as Chat GPT 5.2 and Gemini. The study involved 55 ninth-grade students from a comprehensive school in Almaty. Students' ages ranged from 15 to 16 years. The study participants were assessed on their writing and presentation of an essay on a given topic. The study used a convenience sampling method. Students completed **an essay writing task** entitled *"The Role of Education in Shaping a Successful Future"*. Students completed a minimum of 150 and a maximum of 180 words essay on the assigned topic during a regular English writing lesson. The written essay was then graded by the teacher using Chat GPT 5.2 and Gemini. The teacher was unaware of Large Language Models' grading during the grading process. Participants were assessed according to four criteria: organization, vocabulary, grammar, task achievement. Each criterion was scored on a five-point scale, resulting in a maximum total score of 20 points. All essays were evaluated by three independent raters an experienced English teacher, GPT 5.2 and Gemini. All anonymous essays were first gathered.

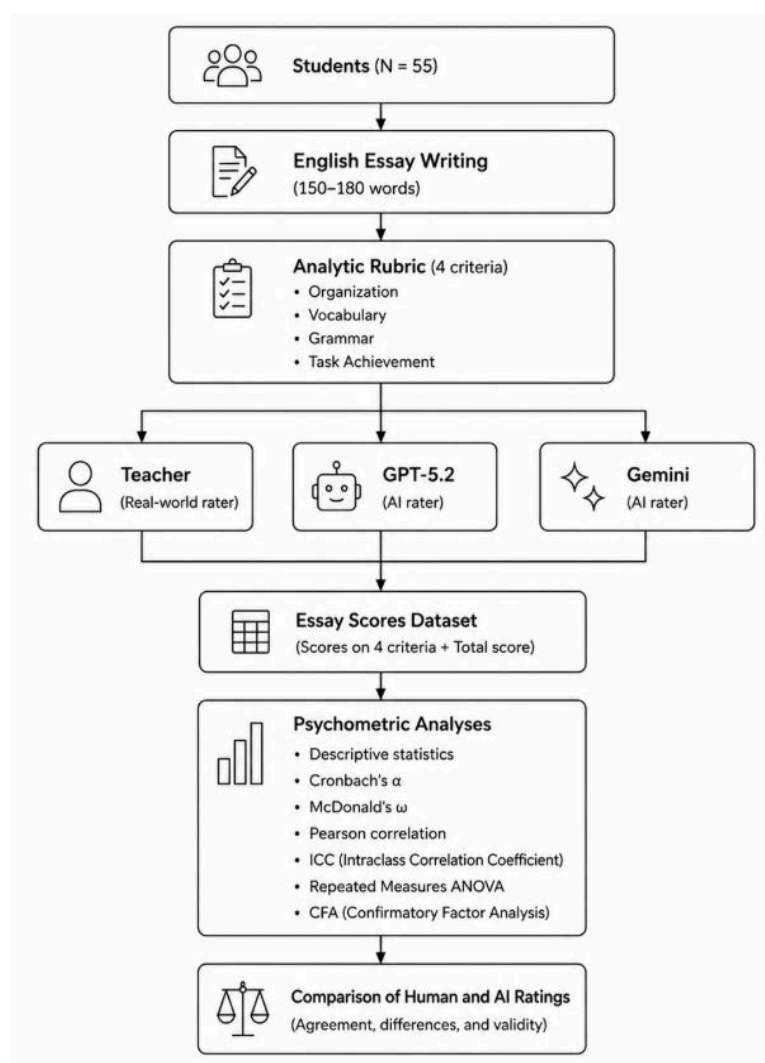


Figure 1 - Workflow of essay evaluation and psychometric assessment procedure

To ensure consistency across AI-based evaluations, the same prompt was used for GPT-5.2 and Gemini. The complete prompt is provided in Appendices A to D. The scores were then imported into Excel and statistical analysis was calculated. Statistical analyses were conducted using Jamovi (version 2.6.45.0). In order to summarize all essay scores descriptive statistics: mean and standard deviation also Cronbach's alpha and McDonald's omega were calculated. Pearson correlation coefficients were calculated to determine relationships among three raters. To examine the construct validity of the writing assessment rubric, Confirmatory Factor Analysis (CFA) was conducted.

Ethical Considerations

Informed consent was obtained from all participants before the study was conducted. Since the participants were school students, permission was also obtained from the school administration by informing their parents. Participants were informed about the purpose of the study and were given full information about their right to withdraw from participation at any time. Participants were informed that their personal information would be kept anonymous and confidential throughout the study.

Results and discussion

Table 1 - Descriptive Statistics of Essay Scores by Rater

Criterion	Teacher M (SD)	GPT-5.2 M (SD)	Gemini M (SD)
Organisation	3.96 (1.03)	3.96 (0.97)	3.78 (1.31)
Vocabulary	3.78 (0.96)	4.17 (0.75)	3.59 (1.10)
Grammar	3.82 (0.96)	3.66 (0.69)	4.14 (0.64)
Content	3.95 (1.02)	3.92 (1.22)	3.94 (0.66)
Total	15.50 (3.71)	15.60 (3.11)	15.40 (3.04)

Table 1 demonstrates the descriptive statistics of the three raters (teacher, GPT-5.2 and Gemini). According to the results, the highest mean score for the vocabulary criterion was shown by GPT-5.2 (M = 4.17, SD = 0.75). The highest mean score for the grammar section was observed in Gemini (M = 4.14, SD = 0.64). The teacher's ratings were relatively similar across the four assessment criteria. As for the overall composite scores, the average results of the three raters were very close to each other (teacher: M = 15.50; GPT-5.2: M = 15.60; Gemini: M = 15.40). These results indicate that the mean scores of the raters did not differ significantly and that their assessment processes were generally consistent.

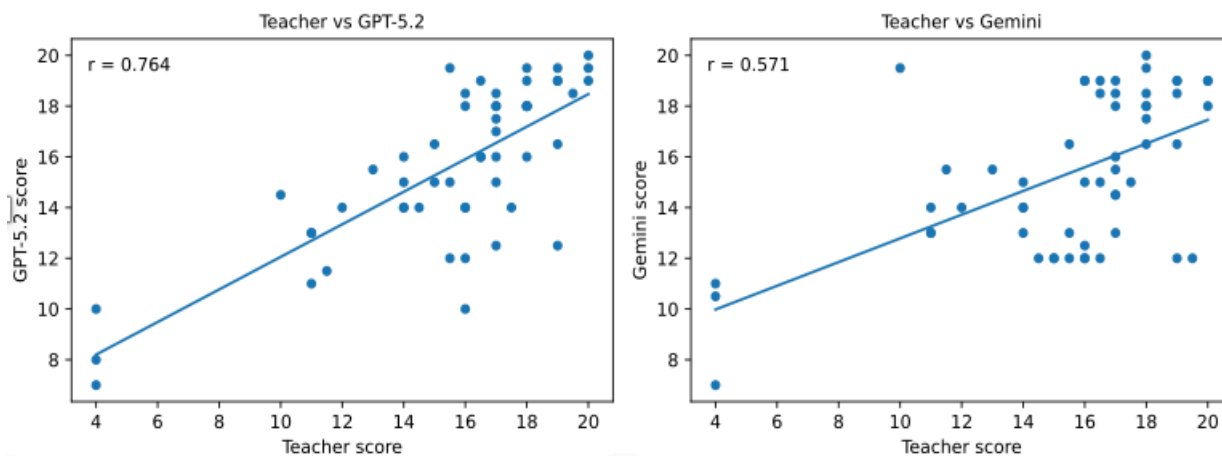
Table 2 - Internal Consistency of the Essay Rating Rubric

Rater	Cronbach's α	McDonald's ω
Teacher	0.947	0.947
GPT-5.2	0.846	0.870
Gemini	0.825	0.857

The highest internal consistency was found in human ratings $\alpha = 0.947$. This suggests that the rubric performed very consistently in human scoring. In addition, GPT-5.2 also used the four criteria consistently, with $\alpha = 0.846$. However, its internal consistency was lower than that of the teacher. Additionally, Gemini demonstrated good reliability (Cronbach's $\alpha = 0.825$). However, its internal consistency is slightly lower compared to GPT and human rater.

Table 3 - Pearson Correlation Matrix Among Essay Ratings

Variable	1	2	3
1. Teacher	-		
2. GPT-5.2	0.764***	-	
3. Gemini	0.571***	0.570***	-

**Figure 2 - Correlation between teacher ratings and AI-generated essay scores (GPT-5.2 and Gemini)**

Pearson correlation analysis was conducted to determine the relationship between human score, GPT-5.2, and Gemini scores. The results revealed a strong positive correlation between teacher and GPT-5.2 scores ($r = .764$, $p < .001$). This indicates that GPT-5.2 scores are highly correlated with teacher scores. Meanwhile, there was a moderate positive correlation between teacher and Gemini ($r = .571$, $p < .001$), as well as GPT-5.2 and Gemini ($r = .570$, $p < .001$). This indicates that Gemini scores are somewhat consistent with teacher scores, but are less correlated than GPT-5.2.

Table 4 - Repeated Measures ANOVA Comparing Essay Scores Across Raters

Source	SS	df	MS	F	<i>p</i>	Partial η^2
Rater	1.61	2	0.806	0.201	.818	.004
Error	432.72	108	4.007	—	—	—

Note. SS = Sum of Squares; MS = Mean Square; Partial η^2 = partial eta squared

A Repeated Measures ANOVA was conducted to determine whether there was a statistically significant difference between the total essay scores of the teacher, GPT-5.2, and Gemini raters. The results of the analysis showed that there was no statistically significant difference between the mean scores of the raters ($F(2,108) = 0.201$, $p = .818$, partial $\eta^2 = .004$). This suggests that the total essay scores of the teacher, GPT-5.2, and Gemini raters were very similar and that the type of rater did not significantly affect the results. Overall, the results obtained indicate that the teacher, GPT-5.2, and Gemini raters rated the essays at a similar level.

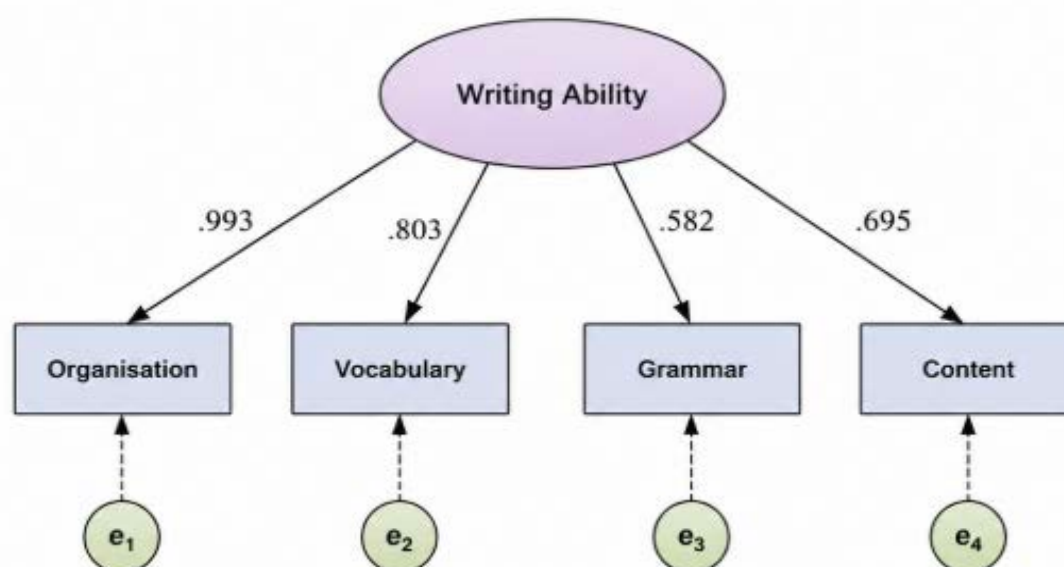
Table 5 - Intraclass Correlation Coefficients (ICC) Between Human and AI Raters

Comparison	ICC	95% CI	F	p
Teacher vs GPT-5.2	0.756	0.615-0.850	7.09	< .001
Teacher vs Gemini	0.563	0.351-0.720	3.54	< .001

In order to investigate inter-rater agreement Interclass Correlation Coefficients (ICC) were calculated for the teacher GPT 5.2 and teacher Gemini comparisons. The results indicate that good agreement between teacher and GPT 5.2 (ICC = .756, 95% CI [.615, .850], $p < .001$). However, teacher and Gemini showed moderate agreement (ICC = .563, 95% CI [.351, .720], $p < .001$). These findings suggest that GPT 5.2 can be more reliable tool for essay scoring.

Table 6 - Confirmatory Factor Analysis Results for the Writing Ability Construct

Indicator	Standardized Loading (λ)	SE	Z	p
Organisation	0.993	0.071	14.06	< .001
Vocabulary	0.803	0.064	12.52	< .001
Grammar	0.582	0.057	10.24	< .001
Content	0.695	0.070	9.88	< .001

**Figure 3 - Confirmatory Factor Analysis Model**

Confirmatory Factor Analysis (CFA) results showed that the four criteria of the writing assessment rubric (Organization, Vocabulary, Grammar, and Content) all measure a common latent construct called “Writing Ability”. All factor loadings were statistically significant ($p < .001$), which proves that these criteria make an important contribution to the assessment of writing ability. Comparing the factor loadings, the Organization indicator had the highest loading ($\lambda = 0.993$), which means that it is the strongest indicator in describing writing ability. It was followed by Vocabulary ($\lambda = 0.803$) and Content ($\lambda = 0.695$). The Grammar indicator showed the lowest factor loading ($\lambda = 0.582$), although its contribution was also statistically significant. In terms of model fit indices, the values of CFI = 0.935 and SRMR = 0.043 indicated an acceptable fit of the model.

However, $TLI = 0.806$ and $RMSEA = 0.262$ this result can be explained by the small sample size in the study ($N = 55$) and the fact that the model consists of only four indicators.

Overall, the study showed that there was no statistically significant difference between the mean scores of the three raters ($p = .818$). This suggests that GPT-5.2 and Gemini systems can produce results close to the teacher's overall grading level. The present findings are consistent with those reported by Seßler et al. [10] and Wu et al. [12] that Large Language Models (LLMs) showed the highest agreement with human raters and good internal consistency. This may be explained by current large language models (LLMs) can produce results that are close to those of teachers when using an analytical assessment rubric.

Furthermore, Chat GPT demonstrated the highest vocabulary scores among all raters. According to results GPT is good at recognizing lexical diversity and word selection, similar to the results reported by Kasneci et al. [4], and García-Varela et al. [9]. This may be attributed that rubric-guided ChatGPT assessment can provide stable and consistent evaluations across analytic writing criteria.

Moreover, Gemini demonstrated higher scores in grammar. A plausible explanation is that large language models are able to effectively learn grammatical structures and language patterns, they can demonstrate high accuracy in detecting grammatical errors. This finding aligns with Mizumoto & Eguchi [5] who reported that Chat GPT has superior ability to detect grammatical errors and assess linguistic accuracy in assessing written work.

Internal consistency results also support this conclusion. Teacher ratings showed the highest reliability ($\alpha = .947$), while GPT-5.2 ($\alpha = .846$) and Gemini ($\alpha = .825$) both had reliability coefficients above .80. This indicates that both AI systems were able to consistently use the analytical rubric. These results are consistent with the findings of a literature review by Lo [6] and a paper published by Kasneci and colleagues [4] These studies show that while modern large language models are able to consistently apply assessment criteria, the professional judgment and assessment stability of experienced teachers still remain at a high level.

The results align with previous research by Abdrasilov et. al [8] who investigated the use of AI assisted Large Language Models for generating tasks for the Unified National Testing (UNT). The research work focused on automatic generation of tasks and their psychometric quality rather than essay scoring. Their findings support human-in-loop approach which is consistent with present study that GPT provides teacher-like results and AI speeds up grading, but human judgment is important because AI does not always fully understand the context. This can be explained by Large Language Models (LLMs) cannot always correctly understand the creative nature of writing, cultural context, or unique responses.

Limitations

First, the limited number of participants ($N = 55$) may have affected the CFA model. High RMSEA in this study indicated that the single-factor model did not fully fit. This may be due primarily to the small sample size and the use of only four indicators. Second, the essays were collected from only one high school and only one task was included in the sample. Third, two large language models, GPT 5.2 and Gemini, were used. Therefore, future studies should increase the sample size and re-examine the structural validity of the model with broader comparison of AI-assisted essay assessment.

Conclusion

Overall, the results of the study showed that GPT-5.2 and Gemini systems can perform assessments that are close to human raters in grading English essays. In particular, GPT-5.2 showed a high level of agreement with teacher ratings. However, it was shown that artificial intelligence systems cannot fully replace the professional judgment of teachers in complex aspects such as understanding students' thinking, assessing creative features, and fully considering the contextual meaning of the text.

In conclusion it is appropriate to consider GPT-5.2 and other large language models not as a substitute for teacher judgment, but as an additional support tool that increases the efficiency and consistency of the assessment process.

Funding information

This research received no external funding.

References

1. Ma M. Exploring learning-oriented assessment in EAP writing classrooms: Teacher and student perspectives // *Language Testing in Asia*. - 2023. - T. 13. - Article 33. - <https://doi.org/10.1186/s40468-023-00249-x>
2. Han T., Zheng Y. AI and human scoring for postgraduate writing: Evaluating score reliability, variability, and rater behaviours // *Studies in Educational Evaluation*. - 2026. - T. 88. - Article 101572. - <https://doi.org/10.1016/j.stueduc.2026.101572>
3. Khamboonruang A. Detecting differential rater severity in a high-stakes EFL classroom writing assessment: A many-facets Rasch measurement approach // *PASAA*. - 2023. - T. 66. - C. 5–36.
4. Kasneci E., Sessler K., Küchemann S., Bannert M., Dementieva D., Fischer F., Gasser U., Groh G., Günnemann S., Hüllermeier E., Krusche S., Kutyniok G., Michaeli T., Nerdel C., Pfeffer J., Poquet O., Sailer M., Schmidt A., Seidel T., Kasneci G. ChatGPT for good? On opportunities and challenges of large language models for education // *Learning and Individual Differences*. - 2023. - T. 103. - Article 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
5. Mizumoto A., Eguchi M. Exploring the potential of using an AI language model for automated essay scoring // *Research Methods in Applied Linguistics*. - 2023. - T. 2, № 2. - Article 100050. - <https://doi.org/10.1016/j.rmal.2023.100050>
6. Lo C. K. What is the impact of ChatGPT on education? A rapid review of the literature // *Education Sciences*. - 2023. - T. 13, № 4. - Article 410. <https://doi.org/10.3390/educsci13040410>
7. Zulpykhar Z., Rakhmetov M., Jugembayeva B., Kapanova D., Zhilmagambetova R. Assessment using artificial intelligence in higher education: Innovations and ethical challenges in the Ibero-American and Kazakh contexts: A mini-review // *Frontiers in Education*. - 2026. - T. 11. - Article 1798126. - <https://doi.org/10.3389/feduc.2026.1798126>
8. Abdrasilov B., Niyazov T., Shinetova L., Altybayeva S., Turalbayeva K., Orlov D. Generation of Kazakhstan's unified national testing variants using AI: A platform for automatic task creation with expert control // *Frontiers in Big Data*. - 2026. - T. 9. - Article 1772101. - <https://doi.org/10.3389/fdata.2026.1772101>
9. García-Varela F., Nussbaum M., Mendoza M., Martínez-Troncoso C., Bekerman Z. ChatGPT as a stable and fair tool for automated essay scoring // *Education Sciences*. - 2025. - T. 15, № 8. - Article 946. - <https://doi.org/10.3390/educsci15080946>
10. Seßler K., Fürstenberg M., Bühler B., Kasneci E. Can AI grade your essays? A comparative analysis of large language models and teacher ratings in multidimensional essay scoring // *Proceedings of the 15th International Learning Analytics and Knowledge Conference*. - New York : Association for Computing Machinery, 2025. - C. 462–472. - <https://doi.org/10.1145/3706468.3706527>
11. Karaçeper R. D., Kiray G. ChatGPT-4o as an automated scoring tool for writing assessment: Strengths and weaknesses // *International Journal of Assessment Tools in Education*. - 2026. - T. 13, № 1. - C. 66–94. - <https://doi.org/10.21449/ijate.1701871>
12. Wu H.-N., Chu M.-N., Hsu J.-L. Comparing GPT and human raters in essay assessment: Variability, bias, and the potential of LLM-based scoring // *Computers and Education Open*. - 2026. - T. 10. - Article 100341. - <https://doi.org/10.1016/j.cao.2026.100341>
13. Yermaganbetova M., Ashimbekova A., Kaibassova D., Akhitova R., Dyusseminina E. Evaluating the impact of an AI-integrated learning platform on student performance: A quasi-experimental study // *Frontiers in Education*. - 2026. - T. 11. - Article 1792353. - <https://doi.org/10.3389/feduc.2026.1792353>
14. East M., Slomp D. The ethical turn in writing assessment: How far have we come, and where do we still need to go? // *Language Teaching*. - 2024. - T. 57, № 2. - C. 262–273. <https://doi.org/10.1017/S0261444823000034>

Ә.Ү. Серікбаева, Д.М. Акижанова

ГЕНЕРАТИВТІ ЖАСАНДЫ ИНТЕЛЛЕКТ АДАМ ЭССЕСІН БАҒАЛАУШЫЛАРДЫ АЛМАСТЫРА АЛА МА? GPT-5.2, GEMINI ЖӘНЕ МҰҒАЛІМ БАҒАЛАУЛАРЫНЫҢ ПСИХОМЕТРИЯЛЫҚ САЛЫСТЫРМАЛЫ ТАЛДАУЫ

Оқушылардың жазба жұмыстарын бағалау маңызды, алайда көп уақытты қажет ететін үдеріс болып табылады. Үлкен тілдік модельдер (LLM) және генеративті жасанды интеллект саласындағы соңғы жетістіктер мұғалімдерге эсселерді бағалауда қолдау көрсетуге мүмкіндік беретін әлеуетті шешімдер ұсынады. Бұл зерттеуде GPT-5.2 және Gemini сияқты үлкен тілдік модельдер адам бағалаушыларымен салыстырылып, олардың сенімділігі психометриялық тұрғыдан бағаланды. Барлығы 55 эссе жинақталып, төрт критерийден тұратын бағалау рубрикасы негізінде GPT-5.2, Gemini және тәжірибелі мұғалім тарапынан бағаланды. Зерттеудің мақсаты – ағылшын тіліндегі эсселерді адам бағалаушыларының, GPT-5.2 және Gemini модельдерінің бағалау нәтижелерін психометриялық тұрғыдан салыстыру, олардың бағалау сенімділігін, өзара байланысын, бағалаудағы айырмашылықтарын және бағалау рубрикасының конструктілік валидтілігін анықтау. Нәтижелер GPT-5.2 сөздік қор критерийі бойынша ең жоғары ұпайларды, ал Gemini грамматика критерийі бойынша ең жоғары ұпайларды бергенін көрсетті. Жалпы алғанда, GPT-5.2 мұғалім бағалауларымен жоғары деңгейде сәйкестік көрсетті және эсселерді бағалауда сенімді қосымша құрал ретінде пайдаланылуы мүмкін. Дегенмен, күрделі мазмұн мен контекстік ерекшеліктерді бағалау кезінде мұғалімнің кәсіби пікірі маңызды болып қала береді. Алынған нәтижелер генеративті жасанды интеллектіні ағылшын тілін шет тілі ретінде оқыту (EFL) контекстінде эсселерді бағалаудың тиімділігі мен бірізділігін арттыратын қосымша құрал ретінде пайдаланудың орынды екенін көрсетеді.

Түйін сөздер: үлкен тілдік модельдер (LLM), GPT-5.2, Gemini, эссені автоматтандырылған бағалау, адамдық бағалаушы, психометриялық бағалау, бағалаушылар арасындағы келісім

Ә.Ү. Серікбаева, Д.М. Акижанова

МОЖЕТ ЛИ ГЕНЕРАТИВНЫЙ ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ ЗАМЕНИТЬ ЭКСПЕРТОВ-ОЦЕНЩИКОВ ЭССЕ? ПСИХОМЕТРИЧЕСКОЕ СРАВНЕНИЕ ОЦЕНОК GPT-5.2, GEMINI И ПРЕПОДАВАТЕЛЕЙ

Оценивание письменных работ учащихся является важным, но в то же время трудоемким процессом. Большие языковые модели (БЯМ) и другие современные достижения в области генеративного искусственного интеллекта предлагают потенциальные решения для оказания поддержки учителям при оценивании эссе. В данном исследовании проводится психометрическая оценка больших языковых моделей, таких как GPT-5.2 и Gemini, в сравнении с экспертами-оценщиками с целью определения надёжности данных решений. Всего было собрано 55 эссе, которые были оценены GPT-5.2, Gemini и опытным учителем с использованием оценочной рубрики, включающей четыре критерия. Цель данного исследования заключается в психометрическом сравнении результатов оценивания английских эссе экспертами-оценщиками, GPT-5.2 и Gemini, а также в определении надёжности оценивания, взаимосвязи между оценками, различий в оценивании и конструктивной валидности оценочной рубрики. Результаты показали, что GPT-5.2 присваивал наиболее высокие баллы по критерию «Лексика», в то время как Gemini демонстрировал наиболее высокие показатели по критерию «Грамматика». В целом GPT-5.2 показал высокий уровень согласованности с оценками учителя и может использоваться в качестве надёжного вспомогательного инструмента при оценивании эссе. Однако профессиональное суждение учителя остается важным при оценивании сложного содержания и контекстуальных особенностей. Полученные результаты подтверждают целесообразность использования генеративного искусственного интеллекта в качестве вспомогательного инструмента для повышения эффективности и согласованности оценивания эссе в контексте преподавания английского языка как иностранного (EFL).

Ключевые слова: автоматизированное оценивание эссе, большие языковые модели (БЯМ), GPT-5.2, Gemini, эксперт-оценщик, оценивание, психометрическая оценка.

References

1. Ma, M. (2023). Exploring learning-oriented assessment in EAP writing classrooms: Teacher and student perspectives. *Language Testing in Asia*, 13, Article 33. <https://doi.org/10.1186/s40468-023-00249-x>
2. Han, T., & Zheng, Y. (2026). AI and human scoring for postgraduate writing: Evaluating score reliability, variability, and rater behaviours. *Studies in Educational Evaluation*, 88, Article 101572. <https://doi.org/10.1016/j.stueduc.2026.101572>
3. Khamboonruang, A. (2023). Detecting differential rater severity in a high-stakes EFL classroom writing assessment: A many-facets Rasch measurement approach. *PASAA*, 66, 5–36.
4. Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, Article 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
5. Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2), Article 100050. <https://doi.org/10.1016/j.rmal.2023.100050>
6. Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences*, 13(4), Article 410. <https://doi.org/10.3390/educsci13040410>
7. Zulpykhar, Z., Rakhmetov, M., Jugembayeva, B., Kapanova, D., & Zhilmagambetova, R. (2026). Assessment using artificial intelligence in higher education: Innovations and ethical challenges in the Ibero-American and Kazakh contexts: A mini-review. *Frontiers in Education*, 11, Article 1798126. <https://doi.org/10.3389/feduc.2026.1798126>
8. Abdrasilov, B., Niyazov, T., Shinetova, L., Altybayeva, S., Turalbayeva, K., & Orlov, D. (2026). Generation of Kazakhstan's unified national testing variants using AI: A platform for automatic task creation with expert control. *Frontiers in Big Data*, 9, Article 1772101. <https://doi.org/10.3389/fdata.2026.1772101>
9. García-Varela, F., Nussbaum, M., Mendoza, M., Martínez-Troncoso, C., & Bekerman, Z. (2025). ChatGPT as a stable and fair tool for automated essay scoring. *Education Sciences*, 15(8), Article 946. <https://doi.org/10.3390/educsci15080946>
10. Seßler, K., Fürstenberg, M., Bühler, B., & Kasneci, E. (2025). Can AI grade your essays? A comparative analysis of large language models and teacher ratings in multidimensional essay scoring. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference* (pp. 462–472). Association for Computing Machinery. <https://doi.org/10.1145/3706468.3706527>
11. Karaçeper, R. D., & Kiray, G. (2026). ChatGPT-4o as an automated scoring tool for writing assessment: Strengths and weaknesses. *International Journal of Assessment Tools in Education*, 13(1), 66–94. <https://doi.org/10.21449/ijate.1701871>
12. Wu, H.-N., Chu, M.-N., & Hsu, J.-L. (2026). Comparing GPT and human raters in essay assessment: Variability, bias, and the potential of LLM-based scoring. *Computers and Education Open*, 10, Article 100341. <https://doi.org/10.1016/j.cao.2026.100341>
13. Yermaganbetova, M., Ashimbekova, A., Kaibassova, D., Akhitova, R., & Dyusseimbina, E. (2026). Evaluating the impact of an AI-integrated learning platform on student performance: A quasi-experimental study. *Frontiers in Education*, 11, Article 1792353. <https://doi.org/10.3389/feduc.2026.1792353>
14. East, M., & Slomp, D. (2024). The ethical turn in writing assessment: How far have we come, and where do we still need to go? *Language Teaching*, 57(2), 262–273. <https://doi.org/10.1017/S0261444823000034>

APPENDICES

Appendix A. Student Essay Writing Task

Essay Topic: The Role of Education in Shaping a Successful Future

Instructions

- Word limit: **150–180 words**
- Introduction (30–40 words)
- Body paragraph 1 (45–55 words)
- Body paragraph 2 (45–55 words)
- Conclusion (20–30 words)

Prompt Used for GPT-5.2 and Gemini

The following prompt was used identically for GPT-5.2 and Gemini to ensure consistency across AI-based evaluations.

You are an experienced EFL writing examiner. Evaluate the following student essay using the analytic writing rubric below.

Criteria:

1. Organization (1–5)
2. Vocabulary (1–5)
3. Grammar (1–5)
4. Task Achievement (1–5)

Instructions:

- Read the essay carefully.
- Assign a score from 1 to 5 for each criterion.
- Calculate the total score out of 20.
- Provide brief rubric-based feedback for each criterion.
- Do not rewrite or correct the essay.
- Evaluate each essay independently without considering previous responses.

Appendix B. Example of Teacher Assessment

Criterion	Score
Organization	3.5
Vocabulary	3.5
Grammar	3.0
Task Achievement	4.5
Total	14.5/20

Appendix C. Example of GPT-5.2 Evaluation

Criterion	Score	Feedback
Organization	4.0	Clear overall structure with an introduction, body paragraphs, and a conclusion. Some transitions could be improved.
Vocabulary	3.5	Appropriate vocabulary, although several collocations are unnatural and some repetition occurs.
Grammar	3.0	Frequent grammatical errors, particularly subject–verb agreement and article usage, reduce accuracy.
Task Achievement	3.5	The topic is addressed, but the essay is below the required word limit and ideas could be developed further.
Total	14.0/20	

Appendix D. Example of Gemini Evaluation

Criterion	Score	Feedback
Organization	3.5	The essay follows the required structure, but paragraphing and logical flow can be strengthened.
Vocabulary	3.5	A sufficient range of vocabulary is demonstrated, although spelling and word-choice errors are present.
Grammar	3.0	Several grammatical inaccuracies affect fluency and overall language accuracy.
Task Achievement	4.5	The response addresses the task effectively with relevant ideas and an appropriate example, although the conclusion is rather brief.
Total	14.5/20	

Авторлар туралы мәліметтер:

Серікбаева Әсия Үмбетбайқызы (корреспондент автор) - Әл-Фараби атындағы Қазақ ұлттық университеті, Түркітану және тіл теориясы кафедрасының 3-курс докторанты, Қазақстан, Алматы қаласы e-mail: assiyaserikbayeva237@gmail.com

Акижанова Диана Муратовна, PhD докторы, Л.Н.Гумилев атындағы Еуразия ұлттық университетінің «Аударма теориясы мен практикасы» кафедрасының меңгерушісі, Қазақстан, Астана қаласы e-mail: akizhanova_dm@enu.kz

Сведения об авторах:

Серікбаева Әсия Үмбетбаевна (автор-корреспондент) - докторант 3 курса кафедры тюркологии и теории языка Казахского национального университета имени аль-Фараби, Республика Казахстан, Алматы, e-mail: assiyaserikbayeva237@gmail.com

Акижанова Диана Муратовна - доктор PhD, заведующая кафедрой «Теория и практика перевода» Евразийского национального университета имени Л.Н. Гумилева. Республика Казахстан, Астана, e-mail: akizhanova_dm@enu.kz

Information about the authors:

Serikbayeva Assiya Umbetbaykyzy (corresponding author) - 3rd-year PhD Candidate, Department of Turkology and Language Theory, Al-Farabi Kazakh National University Kazakhstan, Almaty, e-mail: assiyaserikbayeva237@gmail.com

Akizhanova Diana Muratovna - PhD Head of the Department of Translation Theory and Practice, L.N. Gumilyov Eurasian National University, Kazakhstan, Astana, e-mail: akizhanova_dm@enu.kz