

Scoring consistency of teachers and artificial intelligence algorithms in written expression assessment: a comparative quantitative study

Ahmet Akçay

Ağrı İbrahim Çeçen University

Kılıç Köçeri

kilic.koceri@adalet.gov.tr

Ağrı İbrahim Çeçen University

Tuğba Tümen

Bitlis Eren University

Serdar Safalı

Ağrı İbrahim Çeçen University

Article

Keywords: artificial intelligence assessment, written text evaluation, analytic rubric, comparative quantitative research, teacher–AI comparison

Posted Date: June 15th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9457938/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Written expression assessment at the secondary school level involves complex interactions among text genre, evaluator type, and student characteristics. The present study examined the degree of agreement, divergence, and patterning between scores produced by four large language model (LLM)-based algorithms, ChatGPT, Grok, Minimax, and Gemini, and those assigned by human teachers when rating texts composed by secondary school students. A total of 122 students (63 female, 59 male) enrolled in grades 6, 7, and 8 participated in the study. Each student produced one text in a randomly assigned genre: folktale, poetry, or biography. All raters, both human teachers and AI algorithms, evaluated the texts using an identical five-dimensional analytic rubric. Inter-rater agreement was quantified via Pearson correlations and intraclass correlation coefficients (ICC), while score differences across evaluator type, text genre, and gender were examined using one-way ANOVA with Tukey post hoc tests and independent-samples t-tests. ChatGPT demonstrated the highest alignment with teacher scores ($r = .82$, $ICC = .84$), whereas Minimax showed the weakest correspondence ($r = .74$). Text genre exerted a moderate but statistically significant effect on scores ($\eta^2 = .06$), and no evaluator-by-genre interaction was detected. Gender-related score differences were observed exclusively in teacher evaluations ($d = 0.47$), with AI algorithms producing gender-neutral scoring patterns. These findings indicate that LLM-based algorithms hold promise as formative feedback tools, while also revealing algorithm-specific limitations that preclude their direct substitution for human judgment in high-stakes writing assessment.

Introduction

The systematic evaluation of student writing is a central challenge in secondary education. Effective assessment not only monitors individual development but also informs curriculum design and pedagogical practice¹². However, traditional teacher-based grading is susceptible to well-documented sources of variability: holistic impressions, rater fatigue, and implicit evaluative biases can introduce inconsistency even when standardised scoring guides are employed³. These concerns have sparked a longstanding interest in automated scoring systems as a means of enhancing objectivity and scalability.

Automated Essay Scoring (AES) systems have been deployed in large-scale educational contexts for several decades. Tools such as e-rater and Turnitin rely on surface-level linguistic features and statistical models trained on human-scored corpora⁴⁻⁵. While their predictive validity against holistic human ratings is reasonably well established for English, evidence concerning morphologically rich, agglutinative languages such as Turkish remains sparse. More recently, general-purpose large language models (LLMs) including ChatGPT, Gemini, Grok, and Minimax have demonstrated the capacity to perform nuanced text evaluation that extends beyond surface features to encompass coherence, register, and genre-specific conventions⁶⁻⁷. Unlike their predecessors, these systems were not purpose-built for assessment; their evaluative capacity emerges from broad pre-training on diverse textual corpora.

Despite growing empirical interest, several critical gaps persist. First, comparative investigations involving multiple LLMs evaluated against the same human benchmark, using an identical rubric, within

a single study design are rare. Second, virtually no experimental evidence addresses the scoring behaviour of contemporary LLMs for student-generated texts in Turkish, a language whose agglutinative morphology and syntactic flexibility pose distinct computational challenges. Third, the moderating role of text genre on human–AI agreement has received insufficient attention, notwithstanding evidence that structural predictability varies substantially across genres such as biography, folktale, and poetry^{2,8}. Finally, whether AI-based scorers reproduce or neutralise gender-related evaluator biases documented in teacher assessment research⁹ remains largely unexplored.

The present study addresses these gaps by deploying a comparative quantitative design in which 122 secondary school students’ written texts were scored by two independent teachers and four LLM algorithms using an identical five-dimensional analytic rubric. The study pursues four research questions: (1) To what degree do AI algorithm scores align with teacher scores across text genres? (2) Are there statistically significant score differences attributable to evaluator type or text genre? (3) Do teacher and AI evaluations differ as a function of student gender? (4) Is there a significant evaluator-by-genre interaction in scoring patterns? By providing a simultaneous, multi-algorithm comparison grounded in a psychometrically validated instrument, this study contributes both methodological and applied evidence to the emerging literature on AI-assisted writing assessment.

Results

Descriptive statistics

Descriptive statistics for each evaluator-by-genre cell are presented in Table 1. Overall mean scores ranged from 68.3 (Minimax) to 71.9 (teachers), a difference of 3.6 points on the 0–100 scale. Standard deviations were comparable across evaluators (range: 7.8–9.8), indicating similar score dispersion. Biography received the highest mean scores from all evaluators, whereas poetry received the lowest. Teacher scores ($M = 71.9$) and ChatGPT scores ($M = 71.4$) were nearly identical overall.

Table 1
Descriptive statistics by evaluator and text genre (0–100 scale).

Evaluator	Folktale		Poetry		Biography		Overall M
	M	SD	M	SD	M	SD	
Teacher	72.3	8.4	68.7	9.1	74.6	7.8	71.9
ChatGPT	70.8	7.9	71.4	8.6	72.1	8.2	71.4
Gemini	69.5	8.2	70.1	9.0	71.3	7.9	70.3
Grok	68.7	8.8	69.3	9.4	70.4	8.5	69.5
Minimax	67.2	9.1	68.0	9.8	69.7	8.9	68.3

Note. All scores are on a 0–100 scale. Teacher scores represent the mean of two independent raters. M = mean; SD = standard deviation.

Table 1. Descriptive statistics by evaluator and text genre (0–100 scale).

[Insert Table 1 here]

Inter-rater agreement

Pearson correlation coefficients between all evaluator pairs are reported in Table 2. Teacher scores correlated most strongly with ChatGPT ($r = .82, p < .01$) and least strongly with Minimax ($r = .74, p < .01$). All human–AI correlations were statistically significant and exceeded .70, indicating strong agreement. Inter-algorithm correlations were uniformly higher, ranging from .83 to .88, reflecting greater internal consistency among the LLMs than between any LLM and the human raters. Intraclass correlation coefficients corroborated these patterns: the teacher–ChatGPT ICC of .84 falls within the ‘good’ range as defined by Cicchetti’s criteria¹⁰, while the teacher–Minimax ICC of .75 approaches the lower boundary of acceptability.

Table 2
Pearson correlation matrix across all evaluator pairs.

Evaluator	Teacher	ChatGPT	Gemini	Grok	Minimax
Teacher	—	.82**	.79**	.76**	.74**
ChatGPT		—	.88**	.85**	.83**
Gemini			—	.87**	.84**
Grok				—	.86**
Minimax					—

Note. All correlations are significant at $p < .01$. ** $p < .01$. Dashes indicate self-correlations.

Table 2. Pearson correlation matrix across all evaluator pairs.

[Insert Table 2 here]

Evaluator type and text genre effects

A mixed-design ANOVA examined the main effects of evaluator type and text genre, as well as their interaction, on total scores (Table 3). Evaluator type produced a statistically significant main effect ($F(4, \cdot) = 14.23, p < .001, \eta^2 = .11$), representing a large effect by conventional benchmarks¹¹. Text genre also exerted a significant main effect ($F(2, \cdot) = 8.09, p < .001, \eta^2 = .06$), corresponding to a medium effect size. Critically, the evaluator-by-genre interaction was non-significant ($F(8, \cdot) = 0.56, p = .810, \eta^2 = .01$), indicating that scoring patterns across text genres were broadly consistent regardless of whether the rater was human or algorithmic.

Table 3
Mixed-design ANOVA results for evaluator type and text genre effects on total scores.

Source	SS	df	MS	F	p	η^2
Evaluator type	312.4	4	78.1	14.23	< .001	.11
Text genre	88.7	2	44.4	8.09	< .001	.06
Evaluator × Genre	24.6	8	3.1	0.56	.810	.01
<i>Note.</i> SS = sum of squares; df = degrees of freedom; MS = mean square; η^2 = eta-squared (effect size). Effect size benchmarks: .01 = small, .06 = medium, .14 = large (Cohen, 1988). The evaluator × genre interaction was non-significant.						

Tukey post hoc comparisons (Table 4) revealed that teacher scores did not differ significantly from ChatGPT scores (M difference = 0.50, $p = .723, d = 0.06$). A small but significant difference emerged between teachers and Gemini ($p = .041, d = 0.19$). Moderate and statistically significant differences were observed between teachers and Grok ($p = .003, d = 0.27$) and between teachers and Minimax ($p < .001, d = 0.40$). All effect sizes remained below the threshold for a medium effect ($d = 0.50$), confirming practical equivalence between teacher and AI scores despite reaching statistical significance.

Table 4
Tukey HSD post hoc pairwise comparisons between evaluator groups.

Evaluator (I)	Evaluator (J)	M Difference	p	Cohen's d
Teacher	ChatGPT	+ 0.50	.723	0.06
Teacher	Gemini	+ 1.60	.041*	0.19
Teacher	Grok	+ 2.40	.003**	0.27
Teacher	Minimax	+ 3.60	< .001***	0.40
ChatGPT	Minimax	+ 3.10	< .001***	0.34
Gemini	Grok	+ 0.80	.612	0.09
<i>Note.</i> M Difference = mean score of Evaluator I minus mean score of Evaluator J. Cohen's d = effect size: 0.20 = small, 0.50 = medium, 0.80 = large (Cohen, 1988). Only theoretically meaningful pairwise contrasts are presented. *p < .05. **p < .01. ***p < .001.				

Table 3. Mixed-design ANOVA results for evaluator type and text genre.

[Insert Table 3 here]

Table 4. Tukey post hoc comparisons between evaluator groups.

[Insert Table 4 here]

Gender effects

Independent-samples t-tests compared male and female student scores separately for each evaluator. A significant gender difference was detected only for teacher scores ($t(120) = 2.67$, $p = .009$, $d = 0.47$), with female students receiving higher mean ratings ($M = 73.8$) than male students ($M = 69.9$). None of the four AI algorithms produced a statistically significant gender difference (all $p > .26$), and all associated effect sizes were negligible ($d \leq 0.19$). Full results are presented in Table 5.

Table 5
Independent-samples t-test results for gender differences by evaluator.

Evaluator	Female		Male		<i>t</i>	<i>p</i>	Cohen's <i>d</i>
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>			
Teacher	73.8	7.9	69.9	8.9	2.67	.009**	0.47
ChatGPT	72.1	7.4	70.6	8.4	1.11	.269	0.19
Gemini	70.8	7.8	69.7	8.7	0.78	.437	0.13
Grok	69.9	8.3	69.0	9.3	0.60	.549	0.10
Minimax	68.4	8.8	68.1	9.5	0.19	.850	0.03
Note. M = mean; SD = standard deviation. Cohen's d effect size classification: < 0.20 = negligible, 0.20–0.49 = small, 0.50–0.79 = medium (Cohen, 1988). **p < .01.							

Table 5. Independent-samples t-test results for gender by evaluator.

[Insert Table 5 here]

Discussion

The present study offers three principal contributions to the literature on AI-assisted writing assessment. First, it provides a multi-algorithm, multi-genre comparison conducted within a single psychometrically controlled design a methodological advance over single-system studies that cannot adjudicate algorithm-specific differences. Second, it supplies rare empirical evidence concerning LLM scoring behaviour for an agglutinative language. Third, it introduces gender as an explicit moderator variable, thereby extending prior work that has treated evaluator type and text properties as the sole sources of scoring variance.

The strong alignment between ChatGPT and teacher scores ($r = .82$, $ICC = .84$) is consistent with findings reported by Mizumoto and Eguchi⁶, who observed correlations in the .78–.85 range between ChatGPT and expert raters for English writing samples. The present study extends this result to Turkish, suggesting that the cross-lingual representativeness of ChatGPT's training corpus is sufficient to support a valid assessment even in morphologically complex languages. By contrast, Minimax's weaker agreement ($r = .74$) is consistent with evidence that its architecture was optimised primarily for Chinese-language generation¹², which may limit its sensitivity to the evaluative dimensions privileged in Turkish writing pedagogy.

The non-significant evaluator-by-genre interaction ($\eta^2 = .01$) indicates that AI algorithms did not exhibit genre-specific scoring biases relative to teachers. All evaluators assigned the highest scores to biography texts, likely reflecting that genre's structural predictability and reliance on verifiable content². Poetry received consistently lower scores across all raters, which aligns with theoretical accounts of

poetry's resistance to standardised rubric operationalisation owing to its dependence on aesthetic and figurative dimensions that resist discrete scoring categories¹⁸. These patterns suggest that genre effects on scoring are a property of the texts rather than an artefact of evaluator type.

The gender finding warrants careful interpretation. Teachers awarded female students significantly higher scores ($d = 0.47$), a pattern consistent with prior research documenting implicit evaluator biases in open-ended writing tasks⁹. The absence of analogous gender effects in AI scores might appear to constitute an advantage for algorithmic assessment; however, the neutrality of LLMs should not be conflated with the absence of bias. As Ramesh and Sanampudi⁴ note, biases embedded in pre-training corpora may manifest as differential scoring sensitivity that is not immediately detectable through mean comparisons alone. Future work employing counterfactual fairness analyses is needed to examine whether implicit gender-related patterns are encoded in LLM scoring behaviour.

Taken together, the present findings argue for a differentiated deployment model in which LLMs are incorporated into formative assessment workflows, providing rapid, consistent, and gender-neutral preliminary feedback while final evaluative decisions, particularly those carrying high-stakes consequences, remain the responsibility of trained human educators. This division of labour capitalises on the scalability and consistency of AI systems while preserving the pedagogical judgment and contextual sensitivity that teacher evaluation affords.

Methods

Research design and participants

A comparative quantitative design was employed. The study population comprised students enrolled in grades 6, 7, and 8 at a public secondary school in Turkey during the 2025–2026 academic year. Convenience sampling was used to recruit 122 participants (63 female, 59 male; $M_{\text{age}} = 13.4$ years, $SD = 0.9$). Power analysis following Cohen's¹¹ guidelines confirmed that this sample size provides adequate statistical power ($\beta = .80$) for the detection of medium effect sizes ($f = 0.25$) at $\alpha = .05$. Students were distributed approximately equally across grades and were randomly assigned to one of three writing genres: folktale ($n = 41$), poetry ($n = 40$), and biography ($n = 41$). Ethical approval was granted by the Ethics Committee of Ağrı İbrahim Çeçen University in 2026. All procedures were conducted in accordance with the ethical standards of the institutional research committee and the 1964 Declaration of Helsinki and its subsequent revisions. Informed consent was obtained from parents or legal guardians, and verbal assent was secured from all student participants before data collection.

Writing task and assessment instruments

Students completed the writing task during a single 50-minute class period, producing texts in their own handwriting. Teachers introduced the assigned genre during the preceding week through model texts and guided discussion.

Written expression was assessed using two instruments. First, a holistic teacher rating scale aligned with the national curriculum awarded up to 100 points across five criteria: orthographic and punctuation accuracy (10 points), vocabulary use (20 points), semantic coherence (30 points), structural organisation (15 points), and topical relevance (25 points). Second, an analytic rubric developed by the Turkish Ministry of National Education Measurement, Evaluation, and Examination Services Directorate was applied by both human teachers and AI algorithms. This rubric comprises five dimensions: form (0–20), content (0–25), narrative structure (0–15), language and expression (0–25), and orthography (0–15), yielding a maximum of 100 points. Validity evidence was established via expert review (content validity ratio = .84, based on seven panellists including five academics and two practitioners) and exploratory factor analysis (five-factor solution explaining adequate variance; $\lambda = .62-.81$). Internal consistency was satisfactory ($\alpha = .87$). Inter-rater reliability among human teachers was ICC = .83, and the overall human–AI ICC was .80, both classified as ‘good’ according to Cicchetti’s¹⁰ criteria.

AI evaluation procedure

Student texts were digitised via scanning and transferred to a computer environment. Each of the four LLM-based algorithms ChatGPT (GPT-4o), Gemini (1.5 Pro), Grok (1.5), and Minimax (abab6) received a standardised prompt embedding the full analytic rubric and instructing the model to assign integer scores within each dimension’s specified range. Responses were requested in JSON format to enable systematic extraction of dimensional scores. Teacher and AI evaluations were conducted at separate time points to preclude any influence of AI scores on teacher judgment. Each text was evaluated independently by each algorithm in a separate API call.

Statistical analysis

Normality was assessed using the Kolmogorov–Smirnov statistic and skewness and kurtosis coefficients¹³. Inter-rater agreement was quantified using Pearson correlation coefficients and two-way mixed-model ICCs with absolute agreement. Group differences by evaluator type and text genre were examined using a mixed-design ANOVA, with Tukey’s honestly significant difference test for pairwise comparisons. Gender differences within each evaluator were tested using independent-samples t-tests. Effect sizes were reported as η^2 (ANOVA) and Cohen’s d (t-tests), interpreted according to Cohen’s¹¹ benchmarks. All analyses were conducted using a commercial statistical software package with the significance threshold set at $\alpha = .05$.

Declarations

Competing interests

The authors declare no competing interests.

Funding

This research received no specific funding from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contribution

Kılıç Köçeri and Ahmet Akçay made the primary intellectual and practical contributions to the study, including conceptualization of the research design, development of the methodological framework, data analysis, and drafting of the main manuscript text. Serdar Safalı and Tuğba Tümen contributed to the literature review, data organization, and provided supportive input during the revision process. All authors reviewed and approved the final version of the manuscript.

Acknowledgement

Artificial intelligence (ChatGPT) was used solely for English language editing of this manuscript. All study design, data collection, data analysis, literature review, interpretation of findings, and manuscript writing were performed by the authors.

Data availability

The datasets generated and analysed during the current study are available from the corresponding author on reasonable request.

References

1. Graham, S. & Perin, D. A meta-analysis of writing instruction for adolescent students. *J. Educ. Psychol.* **99**, 445–476 (2007).
2. Myhill, D., Jones, S., Lines, H. & Watson, A. Re-thinking grammar: the impact of embedded grammar teaching on students' writing and metalinguistic understanding. *Res. Pap Educ.* **27**, 139–166 (2012).
3. Wolfe, E. W. Construct validity of automated essay scores. in *Handbook of Automated Essay Evaluation: Current Applications and New Directions* (eds (eds Attali, A. J. & Burstein, K.) 110–122 (Routledge, (2014).
4. Ramesh, D. & Sanampudi, S. K. An automated essay scoring systems: a systematic literature review. *Artif. Intell. Rev.* **55**, 2495–2527 (2022).
5. Knoch, U. Diagnostic writing assessment: the development and validation of a rating scale. *Lang. Test.* **26**, 31–57 (2009).

6. Mizumoto, A. & Eguchi, M. Exploring the potential of using an AI language model for automated essay scoring. *Res. Methods Appl. Linguist.* **2**, 100050 (2023).
7. Dai, W. et al. Can large language models provide feedback to students? A case study on ChatGPT. *Proc. IEEE Int. Conf. Adv. Learn. Technol. (ICALT)* (2023).
8. Weigle, S. C. *Assessing Writing* (Cambridge Univ. Press, 2002).
9. Sadler, P. M. & Hale, A. M. Academic achievement of English language learners in post-NCLB era. *Educ. Policy.* **27**, 575–611 (2013).
10. Cicchetti, D. V. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol. Assess.* **6**, 284–290 (1994).
11. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences* 2nd edn (Lawrence Erlbaum, 1988).
12. Li, J., Reuel, A. & Aggarwal, K. Evaluating language models on multilingual text assessment: implications for cross-linguistic validity. Preprint at (2023). <https://arxiv.org/abs/2311.09783>
13. Creswell, J. W. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches* 4th edn (SAGE, 2014).
14. Warschauer, M. & Grimes, D. Automated writing assessment in the classroom. *Pedagogies* **3**, 22–36 (2008).
15. Baidoo-Anu, D. & Ansah, L. O. Education in the era of generative artificial intelligence (AI): understanding the potential benefits of ChatGPT in promoting teaching and learning. *J. AI.* **7**, 52–62 (2023).
16. Lawshe, C. H. A quantitative approach to content validity. *Pers. Psychol.* **28**, 563–575 (1975).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [ScientificReportsTables.docx](#)