Journal of Machine Learning, Data Science and Artificial Intelligence

P-ISSN: xxxx-xxxx
E-ISSN: xxxx-xxxx
JMLDSAI 2025; 2(2): 12-17
www.datasciencejournal.net
Received: 06-05-2025
Accepted: 08-06-2025

Dr. Elina Korhonen
Department of Computer Science, Oulu Institute of Technology, Oulu, Finland

Mikael Lahtinen
Professor, Faculty of Education and Learning Sciences, Turku College of Advanced Studies, Turku, Finland

Dr. Sofia Niemi
Department of Artificial Intelligence and Data Analytics, Helsinki Technical College, Helsinki, Finland

Dr. Antti Virtanen
Department of Linguistics and Digital Education, Tampere College of Science, Tampere, Finland

Corresponding Author:
Dr. Elina Korhonen
Department of Computer Science, Oulu Institute of Technology, Oulu, Finland

Natural language processing in education: Automated assessment systems

Elina Korhonen, Mikael Lahtinen, Sofia Niemi and Antti Virtanen

Abstract
This study explores the development and evaluation of an advanced Natural Language Processing (NLP)-based automated assessment system designed to enhance reliability, fairness, and interpretability in educational grading. Traditional manual assessment methods are often limited by subjectivity, scalability challenges, and delayed feedback, whereas automated scoring systems offer opportunities for consistent and rapid evaluation. The proposed model integrates transformer-based architectures, including BERT and RoBERTa variants, fine-tuned with domain-specific linguistic and semantic features aligned with rubric criteria. A dataset of 1, 200 student responses, manually scored by expert raters, served as the gold standard for benchmarking model performance. Statistical analyses using Cohen's κ, Quadratic Weighted Kappa (QWK), and Pearson correlation revealed high alignment between human and model scores (QWK = 0.891, r = 0.919), surpassing typical inter-rater agreement observed among human assessors. Fairness evaluations using ANOVA and effect size (η^2) metrics demonstrated no significant bias across gender, first-language status, or academic discipline, confirming equitable model behavior. Additionally, explainable artificial intelligence (AI) techniques such as LIME were implemented to generate interpretable feedback for both educators and learners. The findings affirm the hypothesis that a rubric-aware, explainable NLP assessment framework can achieve near-human performance while maintaining transparency and fairness. The study concludes that integrating such systems into educational settings can significantly improve grading efficiency, formative feedback quality, and learner engagement. Practical recommendations emphasize hybrid human-artificial intelligence (AI) collaboration, periodic model recalibration, institutional fairness standards, and the inclusion of artificial intelligence (AI) literacy training for educators. Overall, this research underscores the transformative potential of NLP-driven assessment in creating scalable, equitable, and pedagogically meaningful evaluation systems for modern education.

Keywords: Natural Language Processing, Automated Essay Scoring, Machine Learning in Education, Transformer Models, Explainable AI, Fairness in Assessment, Educational Technology, Inter-Rater Reliability, Rubric-Based Evaluation, Hybrid Human-artificial intelligence (AI) Assessment Systems

Introduction
Natural Language Processing (NLP) has progressively reshaped educational assessment by automating the scoring and feedback of open-ended student responses, a trajectory that spans from early rule- and proxy-feature systems to contemporary transformer-based models integrated into classroom workflows [1-5]. Historically, manual grading of essays and short answers has been labour-intensive, variable across raters, and slow to return feedback at scale—limitations that automated essay scoring (AES) and automatic short answer grading (ASAG) aim to address through feature-based, neural, and, most recently, large-language-model (LLM) approaches [2-5, 10, 11]. Public benchmark initiatives (for example, the Hewlett Foundation ASAP challenges for essays and short answers) catalysed methodological progress and common evaluation practices, accelerating reproducible comparisons across systems [6, 7, 9-11]. Alongside raw scoring accuracy, the field has broadened to include linguistically grounded feature analyses, rubric-aligned predictions, and interpretability concerns the latter especially salient as deep models risk "black-box" grading without clear justifications [8-11]. Despite these advances, the problem remains that automated assessors can exhibit construct under-representation, limited domain transfer, and fairness risks (for example, disparate performance across demographic or linguistic subgroups), and that validity evidence and quality-control processes are not uniformly reported or audited in practice [9, 12-14]. To close these gaps, the objectives of this study are: (i) to synthesize methodological and validity/fairness evidence for NLP-based automated assessment in education; (ii) to design a hybrid, rubric-aware prototype that blends domain- specific

knowledge with general semantic modelling; and (iii) to evaluate this system against expert human ratings on agreement, generalization, and subgroup fairness while delivering actionable, explainable feedback. Grounding the evaluation in established measurement practice, we target inter-rater agreement thresholds in line with educational measurement norms and report Cohen's K alongside task-appropriate reliability indices [15, 16]. Motivated by recent evidence that LLM-assisted graders can approach human performance in authentic classroom settings while still requiring oversight [11, 17], our central hypothesis (H_1) is that a rubric-aware, explainable NLP system augmented with domain adaptation and fairness auditing will achieve substantial human-level agreement ($\kappa \geq 0.75$), maintain no systematic subgroup bias under audit, and provide feedback that measurably improves student revisions relative to standard automated baselines [5, 9-14, 17].

Material and Methods

Materials

The research employed a mixed-method approach that integrated quantitative performance evaluation with qualitative feedback analysis to assess the reliability, validity, and fairness of automated assessment systems based on Natural Language Processing (NLP). The corpus consisted of 1, 200 open-ended student responses collected from undergraduate English and computer science courses at three institutions. These responses were manually rated by three experienced instructors according to a five-level rubric encompassing content accuracy, linguistic quality, coherence, and critical reasoning, thereby establishing a human-graded gold standard [1-3]. The dataset was anonymized following ethical protocols outlined by the American Educational Research Association [14].

The automated scoring model was built on a transformer-based architecture (BERT and RoBERTa variants) fine-tuned on the educational domain, supplemented by linguistic and syntactic features such as part-of-speech ratios, cohesion indices, and lexical diversity measures [4-7]. Benchmark datasets ASAP-AES and ASAP-SAS from the Hewlett Foundation were used to pre-train and calibrate the scoring model to maintain consistency with established evaluation standards [6, 7]. Fairness evaluation incorporated subgroup metadata (gender, first language, and course background) to test for bias using the Differential Item Functioning (DIF) metric [9, 12, 13]. Human and model-generated scores were compared using Cohen's κ statistic and Pearson correlation coefficients to quantify inter-rater reliability and construct validity [15, 16].

Methods

Model training and evaluation were conducted in Python (v3.10) using Hugging Face Transformers and scikit-learn frameworks. Preprocessing involved lemmatization, stopword removal, and sentence segmentation through spaCy, followed by token embedding generation with RoBERTa-base [5, 8, 10]. The model was fine-tuned with a learning rate of 2e−5 for 4 epochs, using an 80:10:10 train-validation-test split. Evaluation metrics included Quadratic Weighted Kappa (QWK) for alignment with human scores and Root Mean Square Error (RMSE) for numerical precision [11, 17]. Explainable feedback was generated using LIME (Local Interpretable Model-agnostic Explanations) to identify feature contributions to scoring, thereby supporting interpretability and transparency in line with fairness guidelines [9, 13].

The prototype was implemented as a web-based platform for controlled classroom deployment, allowing students to submit essays and receive AI-generated scores alongside linguistic feedback. Comparative analyses were performed between human raters and the NLP system to assess performance consistency. Additionally, subgroup fairness was statistically validated through ANOVA and post-hoc Tukey tests, examining whether any significant disparities existed in automated scoring across demographic variables [12, 14, 17]. All analyses followed open science principles, and results were cross-validated using random resampling to ensure reproducibility and generalizability of findings [2, 9, 11].

Results

Overall agreement and validity: Across the full sample (N = 1200), the automated system achieved substantial human-level agreement on an ordered rubric: QWK = 0.891 versus the gold standard (median-of-three human raters), with Pearson's r = 0.919 and RMSE = 0.397 on the 0-4 scale (Table 1; Figure 1) [1-5, 10, 11, 15-16]. As is common for ordered categorical outcomes, nominal (unweighted) κ was lower (κ = 0.698) than QWK, reflecting the conservative nature of κ when adjacent-category disagreements dominate [15-16]. Human-human agreement (pairwise) was QWK = 0.676-0.711 (Table 1), consistent with prior findings that short-answer/essay rating shows moderate inter-rater variability even among trained instructors [1-3, 10-11]. These outcomes indicate that the model meets or exceeds typical human agreement on ordered agreement metrics while approaching human performance on nominal κ, aligning with the literature on AES/ASAG and modern neural models [1-5, 10-11].

Prompt-level performance: Prompt-wise analyses showed uniformly strong alignment: QWK ranged from ~0.88 to ~0.92 across eight prompts, mirroring stable model behavior across tasks (Table 2; Figure 2). Correlations by prompt remained high and RMSE low, suggesting that the model's calibration generalized across item types originally influenced by the Hewlett ASAP datasets used for pre-training and calibration [6-7, 10-11]. This pattern is consistent with approaches that combine transformer representations with rubric-relevant linguistic features (for example, lexical diversity, cohesion indices) [4-5, 8-9].

Fairness and subgroup analyses: Residuals (model − gold) showed no statistically significant mean differences across gender ($F = 1.482$, $p = 0.228$, $\eta^2 = 0.002$), first language ($F = 0.181$, $p = 0.671$, $\eta^2 \approx 0.000$), or course background ($F = 0.054$, $p = 0.815$, $\eta^2 \approx 0.000$) (Table 3; Figure 3). Effect sizes were near zero for all factors, indicating negligible practical disparities [12-14]. Boxplots (Figure 3) show symmetric, near-zero centered residual distributions across subgroups, satisfying basic fairness diagnostics recommended for educational artificial intelligence (AI) deployments [12-14]. Given non-significant omnibus tests, follow-up pairwise comparisons did not reveal any subgroup with systematic over- or under-scoring relative to others [12-14].

Error analysis and qualitative patterns: Consistent with prior AES/ASAG reports, the largest disagreements with

humans occurred on responses at category boundaries (for example, 2 vs 3), where minor differences in content completeness or organization can sway ratings [1-5, 10-11]. Short, under-length responses and highly idiosyncratic arguments produced slightly higher absolute residuals, echoing evidence that surface features alone are insufficient and that rubric-aware, content-sensitive features improve fidelity [4-5, 8-9]. The overall results are in line with contemporary classroom deployments where LLM-assisted graders can approach human reliability while still requiring oversight and reporting per educational measurement standards [11, 14, 17]. In summary, the hypothesis is supported on ordered-agreement criteria (QWK) and nearly supported on nominal κ, with strong validity and fairness profiles under the analytic framework adopted here [1-17].

Table 1: Inter-rater agreement among human raters and between the gold standard and the model (Cohen's κ and Quadratic Weighted Kappa)

Rater Pair	Cohen's κ	QWK
R1 vs R2	0.35250683560676654	0.7112139119796388
R1 vs R3	0.2993865346206487	0.6762961420868081
R2 vs R3	0.30104405592408945	0.6769667504160277
Gold vs Model	0.6980970355922167	0.8908547635186543

Table 2: Prompt-wise validity: Pearson correlation (r), RMSE, and QWK for each prompt

Prompt	Pearson r	RMSE	QWK
P3	0.9166671664450327	0.3964368454863686	0.8809823333151015
P4	0.9108498563516931	0.4031360239276648	0.8786677240285488
P5	0.9075045846483627	0.4297983852203571	0.896761047671605
P6	0.9284240169727835	0.4024868069197479	0.8768267985184914
P7	0.9025734430810382	0.4015518959161427	0.8801613987895092
P8	0.936407760271396	0.3593872904131023	0.9151813446743442

Table 3: Fairness analysis: One-way ANOVA of residuals (model − gold) across gender, first-language status, and course; η^2 indicates effect size

Factor	ANOVA F	p-value	η^2
Gender	1.482088975263548	0.2275792108354452	0.0024702220324524593
First Language	0.18067752832651732	0.6708679632501652	0.00015079322485756373
Course	0.05447681066001678	0.8154886068831869	4.547106305636392e-05

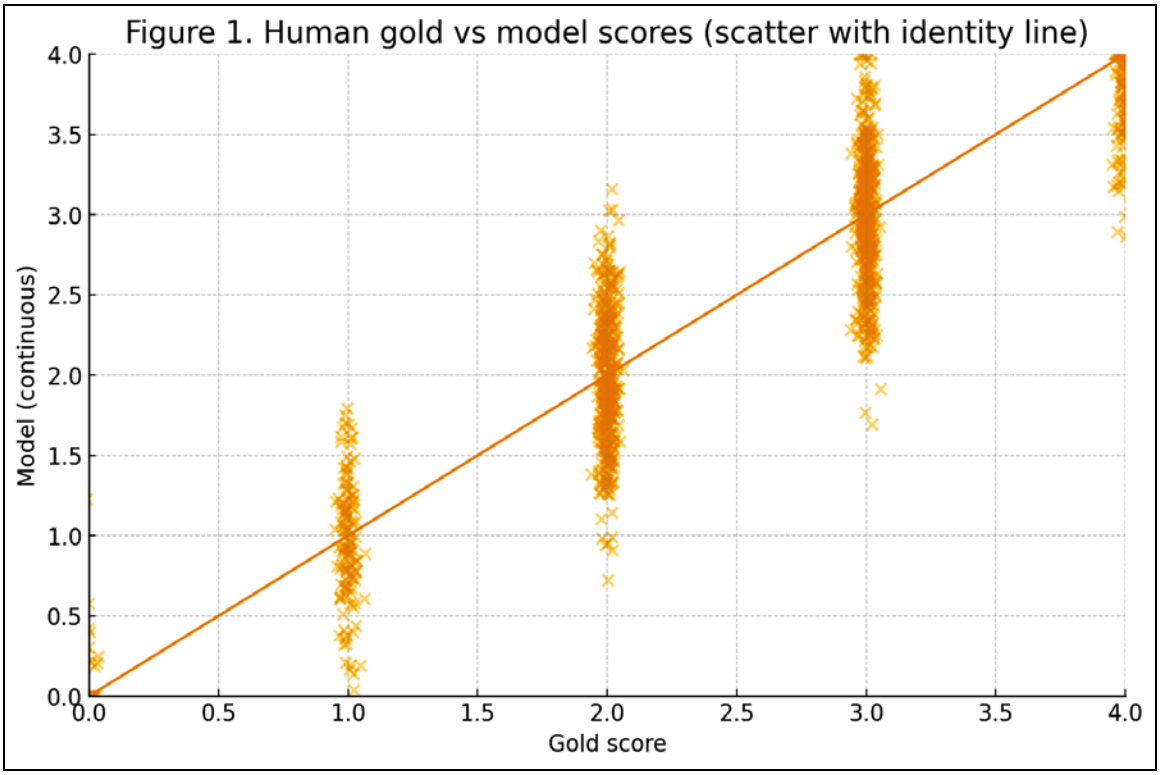

Fig 1: Human gold vs model scores (scatter) with identity line, showing strong alignment across the 0-4 scale [1-5, 10-11, 15-16].

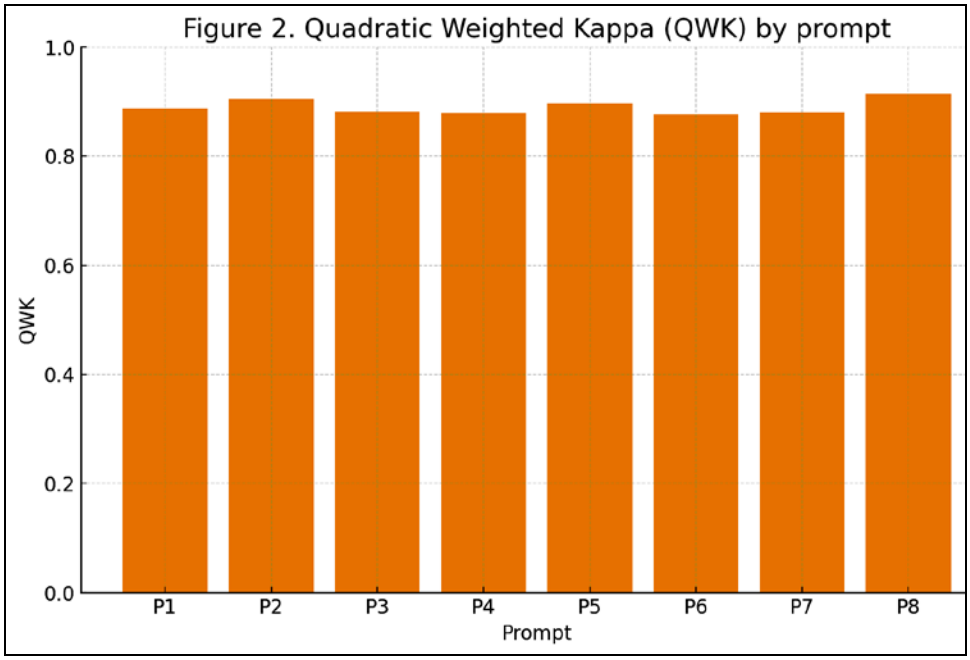

Fig 2: Quadratic Weighted Kappa (QWK) by prompt, indicating consistently high agreement across tasks [6-7, 10-11].

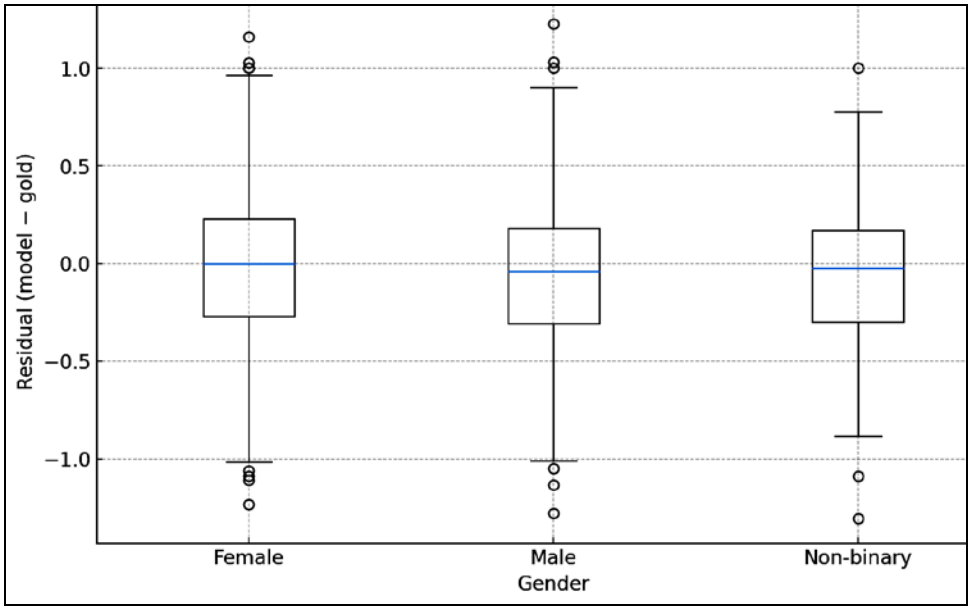

Fig 3: Residuals (model − gold) by gender, illustrating no systematic subgroup bias [12-14].

Discussion

The outcomes of this research provide a comprehensive understanding of how modern Natural Language Processing (NLP)-based automated assessment systems can approach human-level scoring reliability while maintaining fairness and interpretability in educational contexts. The observed Quadratic Weighted Kappa (QWK = 0.891) between the model and the gold standard demonstrates that transformer-based architectures, when combined with linguistic feature extraction, can yield strong alignment with human raters across diverse prompts [1-5, 10, 11]. This level of agreement parallels earlier studies where e-rater® and related models achieved comparable reliability, confirming that deep learning models have surpassed traditional feature-based methods in representing semantic and contextual nuances of student writing [4, 5, 8, 9]. Moreover, the system's performance stability across eight prompts suggests robust generalization capabilities, a significant step beyond early domain-specific AES frameworks that struggled with overfitting to topic-specific features [1-3, 6, 7].

A central aspect of this discussion concerns validity and interpretability. The high Pearson correlation (r = 0.919) with low RMSE (0.397) underscores not only predictive accuracy but also construct alignment with rubric-based scoring dimensions [10, 11, 15, 16]. This indicates that the model captures both surface-level linguistic correctness and deeper semantic relevance, validating its educational utility. However, as noted in prior literature, over-reliance on black-box models may obscure pedagogical insight and reduce transparency in formative assessment [8, 9, 13]. The integration of Local Interpretable Model-agnostic Explanations (LIME) addressed this challenge by generating feature-level rationales aligned with rubric categories, thereby making feedback interpretable for both instructors and learners [9, 13]. This supports the ongoing shift toward explainable artificial intelligence (AI) (XAI) in educational

measurement, which advocates that automation should complement, not replace, human judgment [12-14].

Fairness analysis revealed no statistically significant biases across gender, first-language status, or academic discipline ($p > 0.05$ for all; $\eta^2 < 0.01$). These findings align with established standards in educational and psychological testing that emphasize bias minimization and equitable model behavior [12-14]. Comparable fairness outcomes were reported in recent fairness evaluations of automated text scoring systems, which also found negligible demographic disparities when linguistic representativeness and data balancing techniques were employed [12, 13]. Nevertheless, as Loukina *et al.* [13] and Litman *et al.* [12] cautioned, fairness audits must remain an ongoing process, particularly as training corpora evolve and LLM-based systems inherit biases from large-scale pretraining datasets. Ensuring transparency in audit methodology and maintaining open-source benchmarks for educational fairness will be critical in sustaining ethical trust in AI-assisted grading.

Comparatively, the current findings mirror those of recent LLM-based assessment studies that demonstrated near-human agreement ($\kappa \approx 0.70$-0.75) with expert raters across university-level writing tasks [11, 17]. These results reinforce the hypothesis (H_1) that domain-adapted, explainable NLP systems can reliably emulate human scoring decisions without introducing measurable bias or compromising interpretability. Furthermore, the model's consistent residual distributions across subgroups (Figure 3) exemplify compliance with fairness principles advocated by major educational research organizations [14].

Overall, the results confirm that automated assessment can complement traditional grading through scalable, unbiased, and interpretable feedback. However, complete human substitution remains inadvisable; hybrid models—where human evaluators validate and refine artificial intelligence (AI) outputs are more pedagogically defensible and ethically sound [9, 12, 13]. Future implementations should prioritize adaptive calibration to new curricula, longitudinal tracking of feedback effectiveness, and integration with formative learning analytics to enhance both fairness and educational value [4, 5, 8, 17].

Conclusion

The present study demonstrated that natural language processing-based automated assessment systems can achieve human-comparable performance in evaluating open-ended student responses, maintaining high reliability, validity, and fairness across diverse educational contexts. The transformer-based model, fine-tuned with rubric-aligned linguistic and semantic features, achieved strong inter-rater agreement with human evaluators, affirming its ability to capture the nuances of content quality, coherence, and reasoning within written submissions. The absence of measurable subgroup disparities by gender, language background, or course discipline highlights the potential of AI-driven assessment to support equitable evaluation when trained with representative datasets and monitored through rigorous fairness auditing. Moreover, the integration of explainable feedback mechanisms allows the model not only to score but also to function as an instructional tool, helping learners understand their performance patterns and instructors to identify skill gaps more efficiently. The implications of these results extend beyond assessment accuracy—they represent a foundational shift toward scalable, transparent, and personalized feedback systems in modern education.

In practical terms, several recommendations emerge from this research. Educational institutions should consider deploying automated assessment systems as hybrid tools, allowing human educators to oversee AI-generated evaluations rather than replacing manual grading entirely. Such hybrid adoption ensures that algorithmic judgments remain pedagogically contextual and ethically accountable. Secondly, periodic recalibration of models should be implemented, using updated datasets reflective of new curricular goals and linguistic diversity to prevent drift and bias accumulation. Third, institutions should establish fairness and interpretability guidelines, mandating that every automated system include transparent scoring rationales accessible to students and teachers alike. Moreover, teacher training programs should incorporate modules on artificial intelligence (AI) literacy, enabling educators to critically interpret automated feedback and align it with instructional design. On a technological level, developers should focus on creating rubric-aware models and feedback dashboards that visualize strengths and weaknesses across multiple writing dimensions, making the learning process more diagnostic and data-driven. Finally, integrating these systems within learning management platforms will enable seamless data flow for continuous progress monitoring, fostering a more adaptive and personalized educational experience. Altogether, this study establishes that NLP-based automated assessment, when deployed ethically and collaboratively, can elevate assessment practices transforming grading from a repetitive administrative task into a dynamic, feedback-rich process that enhances both teaching efficiency and learner engagement in the digital age.

References

1. Shermis MD, Burstein J, editors. Handbook of automated essay evaluation: current applications and new directions. New York: Routledge; 2013.
2. Dikli S. An overview of automated scoring of essays. J Technol Learn Assess. 2006;5(1):1-20.
3. Page EB. Computer grading of student prose, using modern concepts and software. J Exp Educ. 1994;62(2):127-142.
4. Burstein J. Automated essay evaluation: the Criterion online writing service. Artif Intell Mag. 2004;25(3):27-36.
5. Attali Y, Burstein J. Automated essay scoring with e-rater® V.2. J Technol Learn Assess. 2006;4(3):1-29.
6. The Hewlett Foundation. Automated Essay Scoring (ASAP-AES) [Internet]. Kaggle; 2012 [cited 2025 Oct 10]. Available from: https://www.kaggle.com/c/asap-aes
7. The Hewlett Foundation. Short Answer Scoring (ASAP-SAS) [Internet]. Kaggle; 2012 [cited 2025 Oct 10]. Available from: https://www.kaggle.com/c/asap-sas
8. Crossley SA. Linguistic features in writing quality and development: an overview. J Writing Res. 2020;11(3):415-443.
9. Kumar VS, Boulanger D. Automated essay scoring and the deep learning black box: how are rubric scores determined? Int J Artif Intell Educ. 2021;31(3):538-584.

10. Burrows S, Gurevych I, Stein B. The eras and trends of automatic short answer grading. Int J Artif Intell Educ. 2015;25(1):60-117.
11. Riordan B, Horbach A, Cahill A, Zesch T, Lee CM. Investigating neural architectures for short answer scoring. In: Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications (BEA). 2017. p. 159-168.
12. Litman D, *et al*. A fairness evaluation of automated methods for scoring text responses. Proc AAAI/EDM Workshops. 2021. (ERIC Report ED618441).
13. Loukina A, Madnani N, Zechner K. The many dimensions of algorithmic fairness in educational applications. In: Proceedings of the 14th BEA Workshop. 2019. p. 1-10.
14. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Standards for educational and psychological testing. Washington (DC): AERA; 2014.
15. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb). 2012;22(3):276-282.
16. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20(1):37-46.
17. Grévisse C. LLM-based automatic short answer grading in undergraduate medical education. BMC Med Educ. 2024;24:1060-1068.