4-бөлім **Раздел 4** **Section 4**

Информатика **Информатика** **Computer science**

IRSTI 28.23.25 DOI: https://doi.org/10.26577/JMMCS209110426

A. Sultanali*, G. Kazbekova
Khoja Akhmet Yassawi International Kazakh-Turkish University, Turkistan, Kazakhstan
*e-mail: aigerim.sultanali@ayu.edu.kz

CROSS-PROMPT TRANSFER AND SUBGROUP ERROR STABILITY IN AUTOMATED ESSAY SCORING

Automated essay scoring (AES) can support large-scale assessment, but strong performance on a random test split does not establish that a model measures writing quality robustly across tasks or produces comparable errors for different student groups. This study evaluates these validity dimensions jointly on the public ASAP 2.0 source-based writing corpus. The analyzed training file contains 24,728 essays from seven prompts with human holistic scores and demographic metadata. A transparent word-count-only Ridge regressor was compared with a histogram-based gradient boosting regressor trained on seven interpretable textual features. Performance was evaluated using quadratic weighted kappa (QWK), mean absolute error (MAE), root mean squared error (RMSE), Spearman correlation, and adjacent agreement. Uncertainty was quantified with 500 bootstrap resamples. Generalization was examined through leave-one-prompt-out validation, supplemented by a prompt-normalized score-scale ablation. Subgroup error differences were tested using Mann–Whitney U tests, rank-biserial effect sizes, and Holm correction. The engineered-feature model achieved QWK=0.732 (95% CI: 0.718–0.746), MAE=0.488, RMSE=0.641, and Spearman ρ=0.788, outperforming the length-only baseline (QWK=0.651). However, word count alone correlated strongly with human score (ρ=0.733, $p < 0.001$). Leave-one-prompt-out QWK ranged from 0.520 to 0.768, and score normalization did not remove the weakest transfer result. After Holm correction, essays from English language learners showed a statistically significant but small increase in absolute error (adjusted $p < 0.001$, $|r| = 0.099$). The study contributes an interpretable evaluation of how strong aggregate scoring agreement can coexist with variable cross-prompt transfer and subgroup-specific error differences, using a transparent length-only baseline and a seven-feature nonlinear model. Practically, the framework can be used to screen AES models before educational deployment and to identify conditions requiring additional human review. The findings indicate that aggregate agreement should be supplemented with task-transfer, uncertainty, and subgroup-error analyses before AES models are considered for educational use.

Keywords: automated essay scoring, machine learning, cross-prompt generalization, fairness, writing assessment.

А. Сұлтанәлі*, Г. Казбекова
Қожа Ахмет Ясауи атындағы Халықаралық қазақ-түрік университеті, Түркістан, Қазақстан
*e-mail: aigerim.sultanali@ayu.edu.kz

Эссені автоматты бағалаудағы тапсырмааралық тасымалдану және топтар бойынша қателердің тұрақтылығы

Жазбаша жұмыстарды автоматты бағалау жүйелері үлкен көлемдегі мәтіндерді тексеруді жеделдетеді, алайда кездейсоқ тест жиынында алынған жоғары дәлдік модельдің жаңа тапсырмаларда тұрақты жұмыс істейтінін немесе әртүрлі білім алушылар топтары үшін қателердің біркелкі болатынын білдірмейді. Зерттеудің мақсаты – ASAP 2.0 ашық корпусында автоматты бағалаудың дәлдігін, мәтін ұзындығына тәуелділігін, жаңа тапсырмаларға жалпылануын және топтық қателердің тұрақтылығын бір эксперименттік хаттамада тексеру. Талдауға жеті тапсырма бойынша 24 728 эссе енгізілді. Мәтін ұзындығына негізделген Ridge регрессиясы базалық әдіс ретінде, ал жеті интерпретацияланатын мәтіндік белгіге негізделген

© 2026 Al-Farabi Kazakh National University

HistGradientBoostingRegressor негізгі модель ретінде қолданылды. Нәтижелер QWK, MAE, RMSE, Spearman корреляциясы және көршілес балл бойынша сәйкестік арқылы бағаланды. Сенімділік 500 bootstrap қайта іріктеуімен анықталды. Жалпылану Leave-One-Prompt-Out хаттамасымен және тапсырма шкаласын қалыпқа келтіру абляциясымен тексерілді. Топтық айырмашылықтар Mann–Whitney U тесті, rank-biserial әсер көлемі және Holm түзетуі арқылы бағаланды. Негізгі модель QWK=0,732 (95% CI: 0,718–0,746), MAE=0,488, RMSE=0,641 және Спирмен корреляциясы $\rho = 0{,}788$ нәтижелерін көрсетті. Бұл модель сөз санына ғана негізделген базалық Ridge моделінен (QWK=0,651) жоғары нәтиже көрсетті. Сөз саны мен адам бағасының байланысы күшті болды (ρ=0,733; $p < 0{,}001$). Leave-One-Prompt-Out тексеруінде QWK 0,520–0,768 аралығында өзгерді. Балл шкаласын қалыпқа келтіру ең әлсіз тапсырмааралық тасымалдану нәтижесін жақсартпады. ELL мәртебесі бойынша абсолюттік қате айырмашылығы Holm түзетуінен кейін статистикалық мәнді болды ($p_{\text{adj}} < 0{,}001$), алайда әсер көлемі шағын болды ($|r| = 0{,}099$). Зерттеудің ғылыми үлесі сөз санына ғана негізделген базалық модель мен жеті мәтіндік белгіге сүйенетін сызықтық емес модельді салыстыру арқылы жоғары жалпы бағалау сәйкестігі жағдайында да тапсырмааралық тасымалдану нәтижелерінің өзгермелілігі мен білім алушылардың жекелеген топтары бойынша бағалау қателерінің айырмашылықтары қатар байқалуы мүмкін екенін түсіндірілетін тәсілмен бағалауында жатыр. Практикалық тұрғыдан бұл тәсіл AES моделін білім беру ортасына енгізер алдында тексеруге және оқытушының қосымша бақылауын қажет ететін жағдайларды анықтауға мүмкіндік береді. Нәтижелер AES жүйелерін тек жалпы дәлдікпен емес, тапсырмааралық тұрақтылық, белгісіздік және топтық қате аудитімен бірге бағалау қажеттігін көрсетеді.

Түйін сөздер: автоматты эссе бағалау, машиналық оқыту, тапсырмааралық жалпылану, әділдік, жазбаша бағалау.

А. Султанали*, Г. Казбекова
Международный казахско-турецкий университет имени Ходжи Ахмеда Ясави, Туркестан, Казахстан
*e-mail: aigerim.sultanali@ayu.edu.kz

Межзаданийный перенос и стабильность ошибок по подгруппам при автоматическом оценивании эссе

Автоматическое оценивание письменных работ позволяет масштабировать проверку больших массивов текстов, однако высокая точность на случайно сформированной тестовой выборке не гарантирует устойчивости модели при переходе к новым заданиям и равномерности ошибок для разных групп обучающихся. Цель исследования – совместно оценить точность автоматического оценивания, зависимость от длины текста, межзаданийную обобщаемость и стабильность групповых ошибок на открытом корпусе ASAP 2.0. В анализ включены 24 728 эссе по семи заданиям. В качестве прозрачной базовой модели использована Ridge-регрессия только по числу слов, а основная модель HistGradientBoostingRegressor обучалась на семи интерпретируемых текстовых признаках. Качество оценивалось с помощью QWK, MAE, RMSE, корреляции Спирмена и adjacent agreement. Неопределенность оценок определялась по 500 bootstrap-выборкам. Межзаданийная устойчивость проверялась протоколом Leave-One-Prompt-Out и дополнительной абляцией с нормализацией шкалы баллов. Межгрупповые различия анализировались критерием Mann–Whitney U, размером эффекта rank-biserial и поправкой Holm. Основная модель достигла QWK=0,732 (95% ДИ: 0,718–0,746), MAE=0,488, RMSE=0,641 и корреляции Спирмена $\rho = 0{,}788$, превзойдя базовую модель Ridge, использующую только количество слов (QWK=0,651). Число слов было сильно связано с экспертной оценкой (ρ=0,733; $p < 0{,}001$). В Leave-One-Prompt-Out проверке QWK изменялся от 0,520 до 0,768. Нормализация шкалы баллов не устранила наиболее слабый результат межзаданийного переноса. Для статуса ELL различие абсолютной ошибки сохранило статистическую значимость после поправки Холма ($p_{\text{adj}} < 0{,}001$), однако размер эффекта был мал ($|r| = 0{,}099$). Научный вклад исследования заключается в интерпретируемой оценке того, как высокое общее согласие автоматических и экспертных оценок может сочетаться с изменчивостью результатов переноса на новые задания и различиями в ошибках

оценивания между отдельными группами обучающихся, на основе сравнения базовой модели, использующей только количество слов, и нелинейной модели с семью текстовыми признаками. Практическая значимость заключается в возможности использовать данный протокол для предварительной проверки AES перед внедрением в образовательную практику и для выявления случаев, требующих дополнительного контроля преподавателя. Полученные результаты показывают, что валидация AES должна включать не только итоговую точность, но и перенос на новые задания, оценку неопределенности и аудит групповых ошибок.
Ключевые слова: автоматическое оценивание эссе, машинное обучение, межзаданийная обобщаемость, справедливость, оценивание письменных работ.

1 Introduction

Automated essay scoring (AES) has progressed from systems dominated by manually designed linguistic indicators to neural and transformer-based models capable of learning complex representations from student writing [1, 2]. Recent work has also emphasized that predictive agreement alone is not sufficient evidence of validity. Active-learning and test-linking studies likewise show that training design and score comparability are central methodological concerns in automated scoring [3, 4]. Machine scores can exploit proxies that correlate with human ratings without adequately representing the intended writing construct, making transparent validity evidence essential [5]. This concern is particularly relevant when superficial properties such as response length carry substantial predictive signal, a long-standing issue in automated writing evaluation [6].

Machine learning approaches to feature-based text analysis have also been investigated in Kazakhstan. Mussiraliyeva et al. investigated the detection of extremist content in Kazakh-language texts using machine learning and deep learning methods, including oversampling and undersampling strategies to address class imbalance [7]. Sultanali and Kazbekova compared lexical and interpretable stylometric representations for detecting AI-generated content in students' written works [8]. Their experiments also examined performance under unseen topics and changes in training-data composition. Although extremist-content detection, AI-generated text detection, and automated essay scoring address different prediction tasks, the cited studies provide relevant methodological context for feature-based text analysis and model evaluation. The present study applies this methodological perspective to automated essay scoring, with particular attention to unseen-prompt transfer and demographic subgroup errors.

A second challenge is generalization across prompts. Prompt-specific models can perform strongly when training and test essays share the same task, yet performance often decreases when the model is transferred to an unseen prompt. Cross-prompt research has therefore explored trait transfer, prompt mapping, and prompt-generalized representations [2, 9–11]. Li and Ng (2024a) argue that reported cross-prompt results are difficult to compare because evaluation protocols, score transformations, and target-prompt assumptions differ across studies [11]. Consequently, a credible AES evaluation should separate within-distribution accuracy from transfer to genuinely unseen tasks. This distinction also aligns with Li and Ng's (2024b) reflection on the AES literature, which calls for evaluation designs that go beyond benchmark-level score agreement [12].

A third concern is fairness. Demographic differences in model error can remain hidden when only aggregate metrics are reported. Kwako and Ormerod examined gender,

race/ethnicity, English language learner (ELL) status, socioeconomic status, and disability status in an XLNet scorer and found small magnifications of some human-score differences [13]. Schaller et al. similarly showed that fairness outcomes depend on algorithmic and data conditions [14]. More recently, Holmes et al. demonstrated that demographic composition of fine-tuning data can affect subgroup bias in LLM-based scoring [15]. Yang et al. evaluated nine automated essay scoring methods using seven metrics and highlighted the potential of traditional machine learning approaches with carefully engineered features to balance accuracy, fairness, and generalizability, although these objectives do not necessarily improve simultaneously [16].

The release of ASAP 2.0 provides an appropriate benchmark for studying these issues because it contains source-based essays, human holistic scores, source materials, and demographic metadata [17]. The present study does not propose a new state-of-the-art semantic scorer. Instead, it uses an interpretable feature-based model to ask a validation-oriented question: *how much confidence should be placed in a seemingly adequate aggregate AES score when length dependence, unseen-prompt transfer, uncertainty, and subgroup error stability are examined explicitly?*

More recently, Cho and Kim (2026) investigated cross-prompt generalization and uncertainty-based human-review selection using 24,728 essays from seven prompts in ASAP 2.0 [18]. Their study used a multiclass logistic classification model and normalized predictive entropy to prioritize essays for review, focusing on major scoring errors and review efficiency. The present study addresses a different but complementary question: how well does an interpretable feature-based scoring model retain its performance across unseen prompts, and what subgroup error differences remain despite acceptable aggregate agreement?

The study makes four contributions. First, it quantifies the predictive role of essay length and compares a word-count-only baseline with a nonlinear model using seven interpretable textual features. Second, it reports uncertainty for the main performance metrics through bootstrap confidence intervals. Third, it evaluates generalization using leave-one-prompt-out validation and a prompt-normalized score-scale ablation. Fourth, it audits subgroup error differences with effect sizes and Holm correction rather than relying on unadjusted significance tests. Together, these analyses provide a compact validation framework for AES systems intended to operate across tasks and student populations.

The study addressed three research questions: (RQ1) how much predictive agreement is obtained from essay length alone, and how much is added by a broader set of interpretable textual features; (RQ2) how stable is model performance when the test prompt is completely unseen during training; and (RQ3) do absolute scoring errors differ systematically across the available demographic subgroups?

Three corresponding hypotheses were evaluated: H1, the seven-feature nonlinear model would outperform the word-count-only baseline; H2, leave-one-prompt-out performance would be lower and more variable than mixed-prompt holdout performance; and H3, subgroup error differences, if present, would be small in effect size after multiplicity correction.

2 Materials and Methods

2.1 Dataset and preprocessing

The experiments used the public ASAP 2.0 source-based writing corpus [17]. The supplied training file contained 24,728 essays from seven prompts. The distribution of essays across the seven prompts is presented in Table 1. Each record included the essay text, a human holistic score, prompt information, source texts, and demographic metadata. The seven prompts were: Driverless Cars (6,170 essays), Facial Action Coding System (4,883), Exploring Venus (4,480), The Face on Mars (3,015), A Cowboy Who Rode the Waves (2,175), Does the Electoral College Work? (2,046), and Car-Free Cities (1,959). Human scores were ordinal; most prompts used a 1–6 scale, whereas one prompt used a 1–5 scale.

Text preprocessing was deliberately minimal to preserve interpretable surface characteristics. Seven features were extracted from each essay: word count, sentence count, mean word length, lexical diversity, punctuation count, digit count, and paragraph count. Words were identified with the Unicode regular expression `\b\w+\b`; sentences were counted after splitting on one or more period, question-mark, or exclamation-mark characters; lexical diversity was the number of unique lower-cased word tokens divided by word count; punctuation and digit counts were character-level counts; and paragraphs were non-empty blocks separated by blank lines. Empty sentence or paragraph counts were floored at one. No demographic attribute was used as a predictive feature. Missing demographic values were retained as missing and excluded only from the corresponding subgroup comparison; therefore, subgroup sample sizes differ by attribute.

Table 1: Distribution of essays across prompts

Prompt	N
Driverless cars	6170
Facial action coding system	4883
Exploring Venus	4480
The Face on Mars	3015
A Cowboy Who Rode the Waves	2175
Does the electoral college work?	2046
Car-free cities	1959
Total	24728

2.2 Study stages

The study proceeded in four stages. First, the corpus was characterized and seven interpretable textual features were extracted. Second, a mixed-prompt holdout experiment compared the length-only baseline with the engineered-feature model and quantified uncertainty by bootstrap resampling. Third, seven leave-one-prompt-out runs and a prompt-normalized score-scale ablation evaluated task transfer. Fourth, subgroup errors were audited with nonparametric tests, effect sizes, and Holm correction. The outputs of these stages were integrated into the Results and Discussion section to evaluate accuracy, robustness, and subgroup error stability jointly.

2.3 Experimental design and models

Two intentionally interpretable regressors were evaluated. The baseline was Ridge regression using only word count. Its purpose was not to maximize performance but to quantify how far a trivial length signal could reproduce human scores. The main model was `HistGradientBoostingRegressor`, trained on all seven textual features. The Ridge penalty was fixed at $\alpha = 1.0$. The HGBR used squared-error loss, learning rate 0.1, a maximum of 100 boosting iterations, 31 maximum leaf nodes, minimum 20 samples per leaf, and zero L2 regularization. These hyperparameter values correspond to the standard default settings of the scikit-learn HistGradientBoostingRegressor. We retained this configuration to provide a transparent and reproducible reference model rather than optimizing hyperparameters for benchmark-specific performance. A fixed random seed (42) was used throughout. Predictions were clipped to the valid prompt score range and rounded to the nearest integer only for QWK and adjacent-agreement calculations; MAE, RMSE, and Spearman correlation used the clipped continuous predictions.

For the mixed-prompt holdout experiment, 80% of essays were assigned to training and 20% to testing using `train_test_split` with `random_state=42` and stratification by the human score. The same partition was used for both models. The central robustness experiment used leave-one-prompt-out (LOPO) validation: in each of seven runs, one prompt was held out entirely for testing and the remaining six prompts were used for training. This design prevents essays from the target prompt from entering model training and therefore provides a direct test of task transfer.

Because one prompt used a different score range, an additional score-scale ablation was performed. Within each prompt, human scores were normalized to the prompt-specific range, the model was trained on normalized targets, and predictions were mapped back to the held-out prompt scale for comparison with the original raw-score LOPO results. This ablation tests whether weak transfer is merely an artifact of incompatible score ranges.

2.4 Evaluation metrics and statistical analysis

The primary metric was quadratic weighted kappa (QWK), which accounts for the ordinal distance between predicted and human scores. Mean absolute error (MAE), root mean squared error (RMSE), Spearman rank correlation, and adjacent agreement (prediction within one score point of the human rating) were also reported. For the main holdout model, 95% confidence intervals were estimated from 500 bootstrap resamples of the test predictions.

For each held-out prompt in the LOPO evaluation, 95% percentile bootstrap confidence intervals for QWK were estimated using 500 paired resamples of the human-assigned scores and the corresponding model predictions. Resampling was performed with replacement within each held-out prompt, using a fixed random seed of 42. The 2.5th and 97.5th percentiles of the bootstrap QWK distribution were used as the interval bounds. Models were not retrained during resampling; therefore, the intervals quantify sampling uncertainty conditional on the fitted models.

Subgroup auditing was conducted for gender, ELL status, economic disadvantage, and disability status. For each group, MAE, mean signed error, and QWK were computed. Pairwise differences in absolute error were tested using two-sided Mann–Whitney U tests. Rank-biserial correlation was reported as an effect-size measure, and Holm correction was

applied across the four demographic comparisons. This combination was chosen because the large sample size can make small distributional differences statistically significant even when their practical magnitude is limited.

3 Results and Discussion

3.1 Score distribution and length-related signal

As shown in Figure 1, human scores were concentrated in the middle of the ordinal scale. Score 3 was most frequent (9,021 essays), followed by score 2 (6,847) and score 4 (5,553); only 200 essays received score 6. The mean essay length was 367.6 words. As illustrated in Figure 2, mean essay length generally increased with higher human score levels: mean word count was approximately 261 for score 1, 256 for score 2, 354 for score 3, 479 for score 4, 634 for score 5, and 779 for score 6. Across all essays, word count correlated strongly with human score (Spearman $\rho = 0.733$, $p < 0.001$).

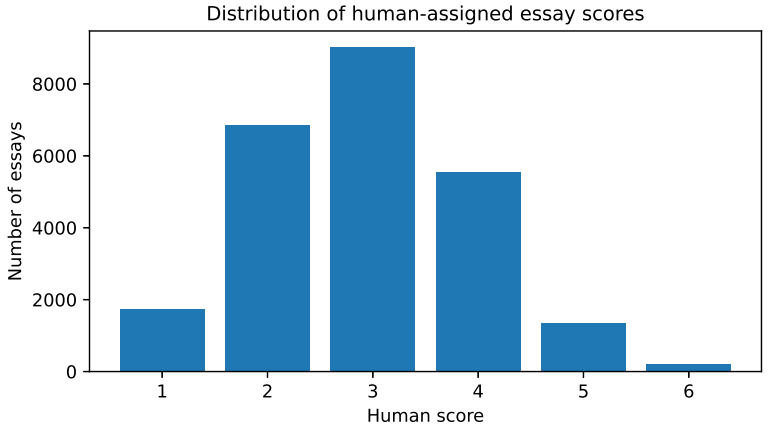


Figure 1: Distribution of human-assigned essay scores.

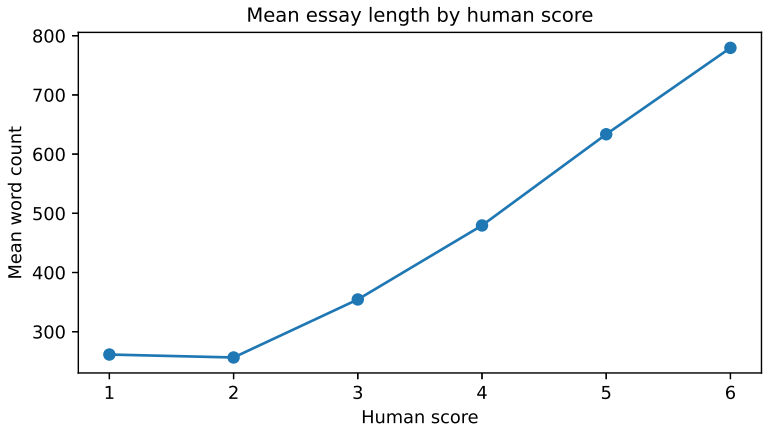


Figure 2: Mean essay length across human score levels.

3.2 Within-distribution scoring performance

The engineered-feature HGBR outperformed the length-only Ridge baseline on every principal metric (Table 2). HGBR achieved QWK=0.732, compared with 0.651 for the baseline, while MAE decreased from 0.555 to 0.488. Figure 3 compares the QWK values of both models together with their 95% bootstrap confidence intervals. Spearman correlation increased from 0.736 to 0.788. Nevertheless, the relatively strong QWK of the length-only model confirms that response length alone carries substantial predictive information.

Table 2: Mixed-prompt holdout performance of the evaluated models

Model	MAE	RMSE	Spearman	QWK	Adjacent acc.
Length-only Ridge	0.555	0.724	0.736	0.651	0.959
Engineered-text HGBR	0.488	0.641	0.788	0.732	0.972

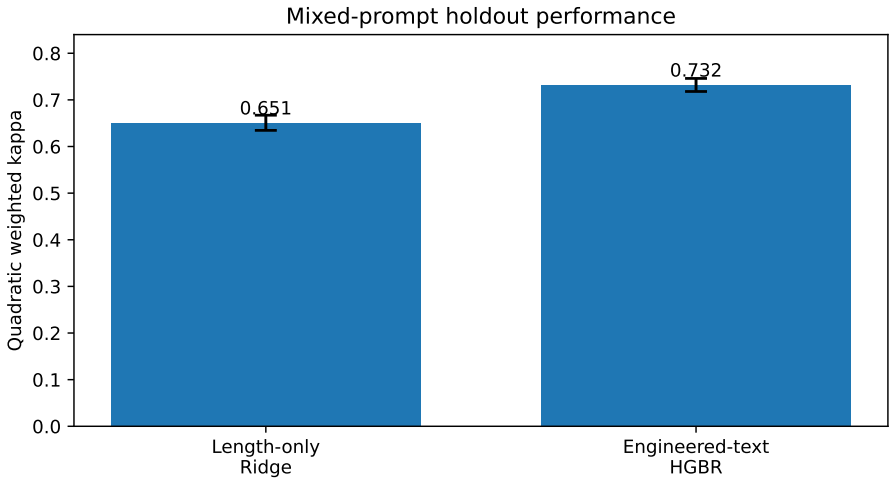


Figure 3: QWK comparison between the length-only Ridge baseline (QWK=0.651, 95% bootstrap CI: 0.635–0.667) and the engineered-text HGBR model (QWK=0.732, 95% bootstrap CI: 0.718–0.746) on the mixed-prompt holdout.

3.3 Generalization to unseen prompts

LOPO evaluation revealed considerably more variation than the mixed-prompt holdout (Table 3). QWK ranged from 0.520 for *A Cowboy Who Rode the Waves* to 0.768 for *Facial Action Coding System*. MAE ranged from 0.444 to 0.600. Thus, the mixed-prompt holdout QWK of 0.732 did not imply uniform transfer to a new task.

Table 3: Leave-one-prompt-out generalization results with 95% bootstrap confidence intervals

Held-out prompt	N	MAE	Spearman	QWK	95% CI
Facial action coding system	4883	0.457	0.815	0.768	0.756–0.780
Driverless cars	6170	0.444	0.815	0.726	0.714–0.737
Exploring Venus	4480	0.516	0.787	0.721	0.707–0.734
The Face on Mars	3015	0.530	0.720	0.649	0.628–0.668
Car-free cities	1959	0.585	0.719	0.660	0.634–0.681
Does the electoral college work?	2046	0.600	0.739	0.652	0.624–0.677
A Cowboy Who Rode the Waves	2175	0.497	0.614	0.520	0.485–0.549

3.4 Prompt-normalized score-scale ablation

The weakest LOPO result occurred on the only prompt scored on a 1–5 rather than 1–6 scale. To test whether this alone explained the reduction, prompt-specific score normalization was introduced. After predictions were rescaled back to the original score ranges, six of seven prompts changed by at most 0.007 QWK. For *A Cowboy Who Rode the Waves*, however, the rescaled QWK decreased from 0.520 to 0.368 rather than improving (Table 4). Figure 4 presents the corresponding QWK values for each held-out prompt without implying an ordered trend across tasks. Thus, the weak transfer cannot be attributed simply to the different score range.

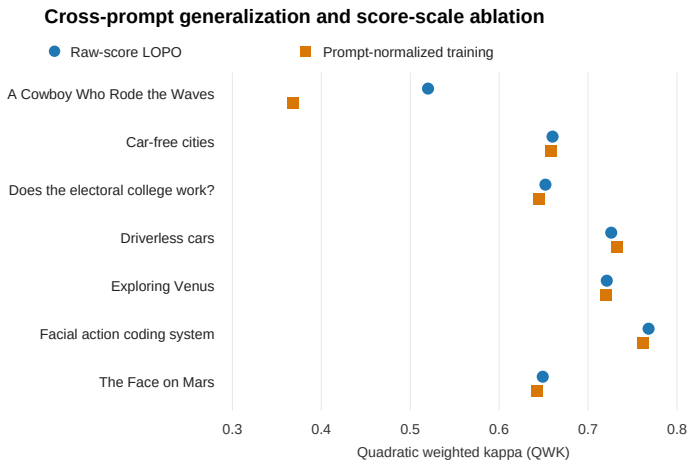


Figure 4: Raw-score LOPO QWK and prompt-normalized score-scale ablation across held-out prompts.

3.5 Subgroup error analysis

Subgroup results are summarized in Table 5. The largest corrected difference occurred for ELL status. Essays from non-ELL students had MAE=0.476, whereas essays from ELL students had MAE=0.567. The Mann–Whitney comparison remained significant after

Table 4: Prompt-normalized score-scale ablation

Held-out prompt	Raw QWK	Rescaled QWK	ΔQWK
A Cowboy Who Rode the Waves	0.520	0.368	−0.152
Car-free cities	0.660	0.658	−0.002
Does the electoral college work?	0.652	0.645	−0.007
Driverless cars	0.726	0.733	+0.007
Exploring Venus	0.721	0.720	−0.001
Facial action coding system	0.768	0.762	−0.006
The Face on Mars	0.649	0.642	−0.007

Holm correction (adjusted $p < 0.001$), but the rank-biserial effect was small in magnitude ($|r| = 0.099$). Gender, economic disadvantage, and disability comparisons were not significant after multiplicity correction.

Table 5: Subgroup performance of the engineered-feature model

Attribute	Group	N	Mean error	MAE	QWK
Gender	Female	2437	0.021	0.499	0.723
	Male	2509	−0.001	0.476	0.736
ELL status	No	4236	−0.022	0.476	0.745
	Yes	626	0.231	0.567	0.618
Economic disadvantage	Disadvantaged	2594	0.028	0.460	0.718
	Not disadvantaged	1531	−0.108	0.471	0.758
Disability status	Identified	605	0.130	0.492	0.651
	Not identified	3520	−0.049	0.459	0.746

3.6 Interpretation of length-related effects

The strength of the association between essay length and human scores is important for validity. A model can appear useful while relying heavily on a surface feature that is only a proxy for writing quality. Perelman previously demonstrated how word-count effects can distort interpretations of automated scoring [6], and Shermis emphasized the need to connect machine-scoring proxies to the intended writing construct [5]. The present findings therefore support the inclusion of an explicit length-only baseline when evaluating automated essay scoring systems.

The engineered-feature model should not be interpreted as a state-of-the-art AES system. Deep neural and transformer-based methods can learn richer semantic representations [1, 19]. The purpose of the engineered model is instead diagnostic: it makes the dependence on observable surface and lexical properties easier to inspect and provides a reproducible reference point for robustness analysis.

3.7 Cross-prompt robustness and transfer limitations

The observed prompt sensitivity is consistent with prior cross-prompt AES research. Prompt mapping and level-aware approaches have been proposed because representations learned on one set of tasks may not transfer reliably to another [9, 10]. Li and Ng (2024a) further caution that cross-prompt AES performance is sensitive to evaluation design and target-prompt assumptions [11]. The present LOPO protocol provides a strict assessment of transfer because no essay from the held-out prompt is available during training.

The score-scale ablation highlights the importance of examining cross-prompt generalization separately from random-split performance. One possible explanation for the observed variation is a prompt-specific change in the relationship between surface-level textual features and human-assigned scores. However, this possibility requires further empirical investigation.

In particular, one plausible explanation for the comparatively weak transfer to the "A Cowboy Who Rode the Waves" prompt is a shift in the distribution of textual features. Because the model relies on seven interpretable lexical and structural indicators, relationships learned from the other six prompts may not generalize equally well to this task. Differences in topic and score distribution may also contribute to the observed performance gap. Moreover, score normalization reduced QWK from 0.520 to 0.368, suggesting that differences in scoring ranges alone cannot explain the weak transfer. These explanations remain hypotheses and require further prompt-level analysis for verification.

The present findings also help qualify the potential of engineered-feature approaches discussed by Yang et al. [16]. Although the seven-feature HGBR model achieved a mixed-prompt QWK of 0.732, its leave-one-prompt-out QWK varied from 0.520 to 0.768. These results show that satisfactory aggregate agreement does not guarantee consistent transfer across unseen prompts. They do not, however, directly contradict the comparative conclusions of Yang et al., since the present study did not evaluate their full set of models under identical conditions.

A related study by Cho and Kim [18] used the same ASAP 2.0 corpus to investigate uncertainty-based selection of essays for human review. Whereas their evaluation emphasized major-error detection and review prioritization, the present analysis focuses on the diagnostic limitations of a transparent word-count baseline and an interpretable nonlinear model, including bootstrap uncertainty, prompt-transfer variability, score-scale normalization, and subgroup-specific scoring errors. The studies therefore address overlapping validity concerns but different evaluation objectives. Their QWK values should not be interpreted as a direct model ranking because the modeling and evaluation procedures differ.

3.8 Subgroup error differences and fairness implications

With a corpus of this size, statistical significance can arise for small distributional differences; effect-size reporting is therefore essential for substantive interpretation. The evidence supports a modest ELL-related performance disparity, rather than establishing that the model is broadly or causally biased against a demographic group. This interpretation is consistent with recent AES fairness research: Kwako and Ormerod reported small demographic magnifications in an XLNet scorer [13], while Schaller et al. showed that fairness depends on model and data conditions [14]. Holmes et al. further found that demographically balanced

fine-tuning data reduced bias relative to demographically restricted training sets [15].

The observed ELL-related error difference may reflect several interacting factors rather than a direct effect of language-learner status. Differences in prompt composition, essay length, and the distribution of surface-level textual features could influence prediction errors, particularly because the model relies on interpretable lexical and structural indicators rather than semantic representations. Although the ELL comparison remained statistically significant after Holm correction, the small rank-biserial effect size suggests that its practical magnitude should be interpreted cautiously. The present observational analysis does not establish whether the disparity arises from model bias, differences in writing characteristics, or other confounding factors. Further prompt-stratified and feature-adjusted analyses are needed to distinguish these possibilities.

3.9 Implications and limitations

Taken together, the experiments show why a single random-split QWK is insufficient as evidence for educational deployment. The engineered model performed reasonably on the mixed-prompt holdout, but three observations qualify that result: length alone was highly predictive; unseen-prompt performance varied widely; and ELL essays showed a small but statistically reliable increase in absolute error. A practical AES validation protocol should therefore report a transparent baseline, uncertainty intervals, task-transfer performance, and subgroup error distributions alongside the headline agreement metric.

The study has several limitations. First, ASAP 2.0 consists of U.S. secondary-school source-based writing, so the findings cannot be directly generalized to university assessment [17]. Second, the model intentionally uses interpretable surface and lexical features and does not claim semantic state-of-the-art performance. Third, demographic metadata contain missing values, so subgroup analyses use available cases. Fourth, the observational design cannot determine why human scores or model errors differ across groups. Finally, future work should compare transformer-based semantic representations under the same LOPO and subgroup-audit protocol and should evaluate whether similar patterns occur in university-level writing.

4 Conclusion

This study evaluated automated essay scoring on 24,728 ASAP 2.0 essays using a validation-oriented, interpretable experimental design. A seven-feature histogram gradient boosting model outperformed a word-count-only baseline, reaching QWK=0.732 (95% CI 0.718–0.746), MAE=0.488, RMSE=0.641, and Spearman $\rho = 0.788$. At the same time, word count correlated strongly with human score ($\rho = 0.733$), demonstrating that length-related signal must be treated as a validity concern rather than ignored. LOPO QWK varied from 0.520 to 0.768, showing that mixed-prompt random testing overstates the stability that can be expected on an unseen task. Prompt-specific score normalization did not explain the weakest transfer result. Subgroup auditing identified a small ELL-related difference in absolute error after Holm correction, while the other tested demographic comparisons were not significant. The central conclusion is methodological: AES should be validated as a system operating across tasks and populations, not only as a predictor optimized for an average benchmark

score. Future work should apply the same validation protocol to transformer-based semantic scorers and to university-level writing in order to test whether the observed transfer and subgroup-error patterns persist in higher-education settings.

Authors contribution

A.S.: conceptualization, methodology, software, formal analysis, investigation, data curation, visualization, writing – original draft preparation. **G.K.:** supervision, methodology, validation, writing – review and editing. All authors have read and approved the final version of the manuscript.

Statement on the use of generative AI

The authors used ChatGPT (OpenAI) for language editing, formatting, bibliography, and technical support in the computational analysis; they reviewed and edited the content and take full responsibility for the publication.

References

[1] Taghipour, Kaveh, and Hwee Tou Ng. 2016. "A Neural Approach to Automated Essay Scoring." In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1882–1891. Austin, TX: Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1193.

[2] Ridley, Robert, Liang He, Xin-yu Dai, Shujian Huang, and Jiajun Chen. 2021. "Automated Cross-Prompt Scoring of Essay Traits." *Proceedings of the AAAI Conference on Artificial Intelligence* 35 (15): 13745–13753. https://doi.org/10.1609/aaai.v35i15.17620.

[3] Firoozi, Tahereh, Hamid Mohammadi, and Mark J. Gierl. 2023. "Using Active Learning Methods to Strategically Select Essays for Automated Scoring." *Educational Measurement: Issues and Practice* 42 (1): 34–43. https://doi.org/10.1111/emip.12537.

[4] Uto, Masaki, and Kota Aramaki. 2024. "Linking Essay-Writing Tests Using Many-Facet Models and Neural Automated Essay Scoring." *Behavior Research Methods* 56: 8450–8479. https://doi.org/10.3758/s13428-024-02485-2.

[5] Shermis, Mark D. 2022. "Anchoring Validity Evidence for Automated Essay Scoring." *Journal of Educational Measurement* 59 (3): 314–337. https://doi.org/10.1111/jedm.12336.

[6] Perelman, Les. 2014. "When 'The State of the Art' Is Counting Words." *Assessing Writing* 21: 104–111. https://doi.org/10.1016/j.asw.2014.05.001.

[7] Mussiraliyeva, Sh. Zh., M. A. Bolatbek, A. N. Zhumakhanova, G. Baispay, and Zh. Medetbek. 2024. "Creating a Model of Semantic Analysis of Extremist Texts in the Kazakh Language." *Journal of Mathematics, Mechanics and Computer Science* 121 (1): 110–121. https://doi.org/10.26577/JMMCS2024121111.

[8] Sultanali, A. E., and G. N. Kazbekova. 2026. "Machine Learning-Based Detection of AI-Generated Content in Students' Written Works." *Academic Scientific Journal of Computer Science*, no. 3: 401–417. https://doi.org/10.32014/2026.2518-1726.470.

[9] Chen, Yuan, and Xia Li. 2023. "PMAES: Prompt-Mapping Contrastive Learning for Cross-Prompt Automated Essay Scoring." In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 1489–1503. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.83.

[10] Chen, Yuan, and Xia Li. 2024. "PLAES: Prompt-Generalized and Level-Aware Learning Framework for Cross-Prompt Automated Essay Scoring." In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 12775–12786. Torino, Italy: ELRA and ICCL. https://aclanthology.org/2024.lrec-main.1118/.

[11] Li, Shengjie, and Vincent Ng. 2024a. "Conundrums in Cross-Prompt Automated Essay Scoring: Making Sense of the State of the Art." In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 7661–7681. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.414.

[12] Li, Shengjie, and Vincent Ng. 2024b. "Automated Essay Scoring: A Reflection on the State of the Art." In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 17876–17888. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.991.

[13] Kwako, Alexander, and Christopher Ormerod. 2024. "Can Language Models Guess Your Identity? Analyzing Demographic Biases in AI Essay Scoring." In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, 78–86. Mexico City, Mexico: Association for Computational Linguistics. https://aclanthology.org/2024.bea-1.7/.

[14] Schaller, Nils-Jonathan, Yuning Ding, Andrea Horbach, Jennifer Meyer, and Thorben Jansen. 2024. "Fairness in Automated Essay Scoring: A Comparative Analysis of Algorithms on German Learner Essays from Secondary Education." In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, 210–221. Association for Computational Linguistics. https://aclanthology.org/2024.bea-1.18/.

[15] Holmes, Langdon, Wesley Morris, Scott Crossley, and Joon Suh Choi. 2026. "Assessing Fairness in Finetuned Scoring Models with Demographically Restricted Training Data." *Assessing Writing* 68 (April): 101032. https://doi.org/10.1016/j.asw.2026.101032.

[16] Yang, Kaixun, Mladen Raković, Yuyang Li, Quanlong Guan, Dragan Gašević, and Guanliang Chen. 2024. "Unveiling the Tapestry of Automated Essay Scoring: A Comprehensive Investigation of Accuracy, Fairness, and Generalizability." *Proceedings of the AAAI Conference on Artificial Intelligence* 38 (20): 22466–22474. https://doi.org/10.1609/aaai.v38i20.30254.

[17] Crossley, Scott A., Perpetual Baffour, L. Burleigh, and Jules King. 2025. "A Large-Scale Corpus for Assessing Source-Based Writing Quality: ASAP 2.0." *Assessing Writing* 65: 100954. https://doi.org/10.1016/j.asw.2025.100954.

[18] Cho, Hyein, and Byeong Jeong Kim. 2026. "Comparison of Review Selection Methods in Automated Essay Scoring: Predictive Uncertainty and Detection of Major Scoring Errors." *Journal of Educational Evaluation* 39 (3): 775–795. https://doi.org/10.31158/JEEV.2026.39.3.775.

[19] Misgna, Haile, Byung-Won On, Ingyu Lee, and Gyu Sang Choi. 2025. "A Survey on Deep Learning-Based Automated Essay Scoring and Feedback Generation." *Artificial Intelligence Review* 58: 36. https://doi.org/10.1007/s10462-024-11017-5.

Information about authors:

Aigerim Sultanali – PhD student at the Khoja Akhmet Yassawi International Kazakh-Turkish University (Turkistan, Kazakhstan, e-mail: aigerim.sultanali@ayu.edu.kz).

Gulnur Kazbekova – Candidate of Technical Sciences, Associate Professor at the Khoja Akhmet Yassawi International Kazakh-Turkish University (Turkistan, Kazakhstan, e-mail: gulnur.kazbekova@ayu.edu.kz).

Авторлар туралы мәлімет:

Сұлтанәлі Айгерім – Қожа Ахмет Ясауи атындағы Халықаралық қазақ-түрік университетінің PhD докторанты (Түркістан, Қазақстан, электрондық пошта: aigerim.sultanali@ayu.edu.kz).

Казбекова Гулнур – техника ғылымдарының кандидаты, Қожа Ахмет Ясауи атындағы Халықаралық қазақ-түрік университетінің қауымдастырылған профессоры (Түркістан, Қазақстан, электрондық пошта: gulnur.kazbekova@ayu.edu.kz).

Сведения об авторах:

Султанали Айгерим – докторант PhD Международного казахско-турецкого университета имени Ходжи Ахмеда Ясави (Туркестан, Казахстан, электронная почта: aigerim.sultanali@ayu.edu.kz).

Казбекова Гулнур – кандидат технических наук, ассоциированный профессор Международного казахско-турецкого университета имени Ходжи Ахмеда Ясави (Туркестан, Казахстан, электронная почта: gulnur.kazbekova@ayu.edu.kz).

Revised: August 26, 2026
Accepted: November 18, 2026