

Rubric-Aligned Generative-AI Features as Supplementary Predictors in Feature-Based Automated Essay Scoring

Yue Huang and Duanli Yan and Corey Palermo

Measurement Incorporated

Correspondence: yueh@udel.edu

Abstract

Feature-based *automated essay scoring* (AES) relies on explicitly defined linguistic features, which are essential but may represent rhetorical qualities like organization and evidence incompletely. Rubric-aligned generative-AI features could fill this gap, yet their incremental value over established linguistic features remains untested. Using 4,572 Grade 7 argumentative essays scored on three traits—Organization/Purpose, Evidence/Elaboration, and Conventions—we compared random forest models using 60 linguistic features, 42 generative-AI features, or a combined feature set. We found that generative-AI features correlated moderately with human trait scores ($|r| = .29$ to $.64$) and overlapped only partially with the linguistic features. On a held-out test set, combined models improved quadratic weighted kappa (QWK) by .011 to .019 over linguistic-only models. Paired bootstrap intervals excluded zero for Organization/Purpose QWK, as well as Conventions exact agreement and mean absolute error. Our findings support the integration of rubric-aligned generative-AI features alongside linguistic features to meaningfully improve trait-based AES.

1 Introduction

Automated essay scoring (AES) uses computational methods to evaluate student written responses and aims to assign valid and reliable scores intended to reflect writing quality (Huang et al., 2025). Its development has been motivated by the need to assess substantial volumes of student writing efficiently and consistently. Comparisons of scoring systems have demonstrated the potential of AES in large-scale assessment, while also highlighting the importance of evaluating automated scores against human judgments (Shermis, 2014). Beyond predictive performance, an important consideration is whether the information underlying a score can be examined and connected to the intended writing

constructs. Feature-based approaches provide a foundation for this examination through explicitly defined linguistic features and documented scoring procedures (Attali and Burstein, 2006).

Recent advances in generative-AI offer opportunities to assess higher-level writing qualities, such as organization, coherence, and idea development, that may be difficult to capture fully through conventional linguistic features. Generative models can respond to rubric instructions and produce judgments concerning multiple aspects of an essay. However, their ability to represent deeper rhetorical qualities accurately remains an empirical question. Huang et al. (2025) identified promising applications in generative-AI scoring alongside continuing gaps in validity evidence. Studies that prompt generative-AI to assign essay scores directly have also reported encouraging agreement with human ratings (e.g., Huang and Wilson, 2025; Huang et al., 2026), particularly for potential low-stakes applications (Tate et al., 2024). These findings motivate examining how generative-AI can contribute to established scoring practices while making its intended judgments available for scrutiny.

The present study investigates generative-AI as a feature-augmentation tool for trait-level scoring of source-based argumentative writing. Structured, rubric-aligned generative-AI features are combined with established linguistic features from an operational AES system in independently fitted scoring models. This approach allows us to examine both their relationships with human trait scores and their additional predictive contribution. From a broader perspective, examining what these different types of features capture may also inform more interpretable scoring and the future development of automated feedback addressing specific aspects of writing quality. The following research questions guided the investigation.

RQ1: How are rubric-aligned generative-AI features associated with human trait scores, and to

what extent do they overlap with established linguistic features within each writing trait?

RQ2: To what extent does adding rubric-aligned generative-AI features improve the performance of linguistic-feature-based scoring models in predicting human trait scores, both overall and within human-score categories?

2 Related Work

Feature-based AES translates writing qualities into measurable textual indicators. Early *Project Essay Grade* (PEG) work combined proxies, or observable approximations of writing characteristics, through multiple regression (Bunch et al., 2017; Page, 1966). Subsequent developments emphasized meaningful feature inventories and expert-informed modeling (Attali and Burstein, 2006). McNamara et al. (2015) combined indices from Coh-Metrix, a tool for analyzing text cohesion and other linguistic chrematistics, the Writing Assessment Tool, and Linguistic Inquiry and Word Count in a hierarchical framework, while Latifi and Gierl (2021) evaluated nine categories of Coh-Metrix features against human ratings of secondary-school essays. These studies illustrate how lexical, syntactic, and discourse information can support scoring beyond surface counts. They also position feature selection as a substantive measurement decision: predictive usefulness depends on how indicators are defined, organized, and modeled, as well as whether their meanings can be connected to the intended writing constructs. Nevertheless, linguistic features may only partially represent context-sensitive judgments about reasoning, idea development, and supporting evidence. Huang et al. (2025) identified deeper rhetorical and pedagogical assessment as a continuing challenge. This motivates examining whether additional features can complement established linguistic features in representing rubric-defined quality. This concern does not diminish the relevance of linguistic features. Instead, it raises the possibility that rubric-guided judgments could represent aspects of writing quality that observable linguistic features capture incompletely.

Hybrid approaches integrate linguistic information with language-model capabilities in different ways. Atkinson and Palma (2025) combined linguistic features, discourse patterns, and GPT (Generative Pre-trained Transformer)-based contextual embeddings, supporting integration across linguis-

tic levels. Hou et al. (2025) instead incorporated extracted linguistic features into large language model (LLM)-based scoring, reporting promising in-domain and out-of-domain results. Rationale-based approaches introduce intermediate assessment information. Chu et al. (2025) generated trait-specific rubric rationales and encoded them alongside essays for score prediction. Do et al. (2025) supplied reference scores to a larger model to generate scoring rationales, then trained a smaller model to produce scores and explanations. Both reported scoring benefits, using rationales as prediction inputs or jointly learned outputs. More directly, Eltanboully et al. (2025) converted trait rubrics into assessment questions with high, medium, or low responses, combining these indicators with generic writing and prompt-specific features in trait regressors. Their positive cross-prompt findings provide a precedent for rubric-based AI feature generation. Together, these studies show how generative-AI can provide rubric-aligned information beyond a final score.

Building on these prior studies, an important question is whether such generative-AI features contribute information relevant to human trait scores beyond that captured by established, theory-informed linguistic features. The present study addresses this question by examining feature associations and overlap alongside the incremental predictive value of combining the two feature sources.

3 Methods

3.1 Sample

The study used a de-identified sample of 4,572 human-scorable essays written by Grade 7 students in response to one source-based argumentative writing item from a large-scale assessment. Responses designated nonscorable by human raters were excluded. Human raters scored the essays using the Smarter Balanced argumentative writing rubric, which specifies Organization/Purpose and Evidence/Elaboration on 1–4 scales and Conventions on a 0–2 scale (Smarter Balanced Assessment Consortium, n.d.).

A randomly selected subset of 1,147 essays (25.1%) received a second human rating. Scores from the first rater (Rater 1) served as the training targets and the reference for model evaluation. Scores from the second rater (Rater 2) were reserved for describing human–human agreement. Using a common 70/30 split, stratified by the Rater

1 Organization/Purpose score, we assigned 3,200 essays to training and 1,372 to testing. The test sample included 350 double-scored essays.

3.2 Linguistic and Generative-AI Features

Two sets of features were obtained for each essay: conventional linguistic features and rubric-aligned generative-AI features. For the conventional linguistic feature set, established indicators were obtained from a contemporary operational version of PEG, a well-established, theory-informed AES system incorporating several hundred *natural language processing* (NLP) features (Bunch et al., 2017). These features span lexical, syntactic, discourse, and conventions categories; individual feature definitions are proprietary and are not reported here. NLP professionals reviewed definitions and groupings of these features, selecting 81 linguistic features and mapping them to the three writing traits. The features retained after preliminary screening were used as predictors in independently fitted scoring models, described in *Section 3.4*. This study only relied on the linguistic feature schemas and did not reproduce the source AES system’s operational scoring procedure.

For the generative-AI features, Claude Haiku 4.5, with temperature set to 0, was used for both codebook generation and essay-level feature scoring. A separate codebook was generated for each trait using its corresponding rubric, a generic description of the source-based argumentative assignment, and grade-level context. Prompts were refined over several rounds by examining the generated-AI features for rubric coverage and relevance to the target trait. The final prompt instructions requested 15 features per trait, each with a unique identifier, feature descriptive name, explicit rubric basis, allowable values, feature scoring anchors, and concise feature scoring guidance. For example, the Organization/Purpose codebook included the feature “Logical progression of ideas”. Allowable values were integers from 1 to 5. The endpoint scoring anchors were “Ideas are randomly ordered or lack clear progression; confusing sequence.” (1) and “Ideas progress logically and strategically; sequence strengthens argument.” (5). Scoring guidance stated: “Evaluate whether paragraphs and ideas follow a coherent order that builds toward the conclusion.” Similarly, “Citation and attribution clarity” (Evidence/Elaboration) anchors from “No citations or attributions; source origin unclear or absent.” (1) to “Clear, consistent, and effective

citations or attributions throughout essay.” (5). For Conventions, “Sentence formation and completeness” anchors from “Frequent fragments, run-ons, or fused sentences throughout.” (1) to “Consistently correct sentence formation throughout.” (5).

All feature ratings ranged from 1 (weakest) to 5 (strongest demonstration of the represented quality), reflecting individual generative-AI features rather than final rubric trait scores. This scale allowed finer differentiation than the three- or four-point trait rubrics and extended Eltanbouly et al.’s (2025) three-category feature ratings to five levels. Rubric-linked definitions and anchors provided an inspectable specification of the generative-AI features’ intended judgments.

The initial generative-AI feature codebooks contained 45 features in total. Human professionals reviewed them for rubric relevance, trait specificity, and within-trait redundancy, removing two Organization/Purpose features and one Evidence/Elaboration feature. The two Organization/Purpose features lacked trait specificity, aligning more closely with Evidence/Elaboration and overlapping with features in that trait. The Evidence/Elaboration feature was removed because of its weaker rubric relevance. Feature scores were subsequently obtained through LLM API interactions. Each interaction included one essay, all three trait codebooks, and instructions directing the model to evaluate each feature according to its definition and scoring anchors.

The essay-writing prompt and source passages were secure and thus not provided to the generative-AI model. Consequently, source-related ratings reflected evidence and attribution apparent within the response, rather than verified accuracy or completeness relative to the source materials. This restriction is particularly relevant to interpreting Evidence/Elaboration features.

3.3 RQ1: Feature Screening and Association Analyses

We performed descriptive screening once on the full training dataset. Within each predefined trait and feature-source group, features were flagged for removal if at least 80% of values were missing, at least 80% were zero, or were perfectly correlated with another feature. We held the retained feature sets fixed for model fitting and evaluation and used the same sets for the test dataset.

For RQ1, we calculated Pearson correlations between each retained feature and its correspond-

ing human trait score in the training sample. We used absolute correlations to compare association strength across features whose scoring directions differed. Within each trait, all generative-AI–linguistic feature pairs were correlated, and their absolute Pearson correlations were summarized.

3.4 RQ2: Scoring Models and Evaluation

We fitted nine focal random forest regression models: three writing traits crossed with linguistic-features-only, generative-AI-features-only, and the combined features. Random forests aggregate predictions across decision trees (Breiman, 2001); we implemented the models with scikit-learn (Pedregosa et al., 2011) in Python 3. All random forest models used 500 trees, no maximum depth restriction, a minimum of two samples per leaf, and a fixed random seed, with all other hyperparameters left at their default values. Thus, the feature conditions were compared under a common model configuration rather than separately optimized hyperparameters. Five-fold cross-validation within the training partition used shuffled folds stratified by the relevant trait score. We then fitted each model to all 3,200 training essays and applied it to the held-out test data. We rounded raw regression predictions to the nearest integer and clipped them to the allowable trait range for agreement analyses. We also calculated mean absolute error (MAE) and R^2 from raw predictions to describe model fit.

For RQ2, quadratic weighted kappa (QWK) served as the primary scoring performance index. Exact agreement, adjacent agreement (i.e., agreement within one score point), and standardized mean difference (SMD) provided complementary information. Human–human agreement metrics were calculated on the smaller subsample with only double-scored test essays and used only as descriptive context. To assess whether adding generative-AI features improved scoring performance, we conducted a paired bootstrap analysis with 10,000 resamples with replacement to estimate 95% percentile confidence intervals for differences in QWK, exact agreement, and MAE (calculated from rounded predictions) between the combined and linguistic-only random forest models. The bootstrap analysis used predictions from the fitted models without refitting.

Finally, score-level analyses examined error distributions within human-score categories. Exact agreement by score point—the proportion as-

signed the same integer score as the human rating—described correct classification within each scoring category. We also compared category mean predictions with the corresponding human scores to summarize the direction and magnitude of average discrepancies.

4 Results

4.1 RQ1: Feature Associations and Overlap

Human review removed three generative-AI features, and descriptive screening removed 21 linguistic features, leaving 42 generative-AI and 60 linguistic features for analysis (Table 1). We examined correlations between each retained feature and its corresponding human trait score and found that generative-AI feature scores were generally positively associated with the corresponding human trait scores. Pearson correlations ranged from .330–.638 for Organization/Purpose, .371–.639 for Evidence/Elaboration, and .289–.506 for Conventions. Linguistic features showed a wider range of association strengths, with absolute Pearson correlations ranging from .014–.740, .002–.743, and .107–.622, for Organization/Purpose, Evidence/Elaboration, and Conventions, respectively.

We ranked all retained features jointly across both feature sets within each trait by their absolute Pearson correlations with human scores for the training dataset. Because linguistic features differed in scoring direction, stronger absolute correlations did not necessarily indicate that higher feature values were associated with better writing. A linguistic feature had the strongest association in every trait, but ranking patterns differed. For Organization/Purpose, the highest-ranked linguistic feature was followed by 12 generative-AI features. For Evidence/Elaboration, linguistic features occupied the two highest positions, followed by three generative-AI features. Linguistic features occupied four of the five highest positions for Conventions. These patterns indicate that multiple generative-AI features were strongly associated with human judgments of Organization/Purpose and Evidence/Elaboration, whereas several established linguistic features showed the strongest associations for Conventions. These individual associations do not establish the relative predictive value of the feature sets, which we evaluated through model comparisons.

Across all within-trait generative-AI–linguistic feature pairs, mean absolute correlations were

Trait	Initial Linguistic Features	Retained Linguistic Features	Initial Generative-AI Features	Retained Generative-AI Features	Combined
Organization/Purpose	25	19	15	13	32
Evidence/Elaboration	39	32	15	14	46
Conventions	17	9	15	15	24
Total	81	60	45	42	102

Table 1: Features retained.

Note. Human review removed 2 Organization/Purpose and 1 Evidence/Elaboration generative-AI features. Descriptive screening on the training data removed 21 linguistic features and no generative-AI features.

.176 for Organization/Purpose, .229 for Evidence/Elaboration, and .262 for Conventions (Table 2). This indicates incomplete pairwise overlap between the two feature sources. However, maximum correlations exceeded .60 in every trait, showing that relatively strong overlap coexisted with the lower average associations. These descriptive patterns are consistent with a mixture of shared and differing information between linguistic features and generative-AI features.

Trait	Feature Pairs	Mean $ r $	Min. $ r $	Max. $ r $
Organization/Purpose	247	.176	.0054	.620
Evidence/Elaboration	448	.229	.0002	.667
Conventions	135	.262	.0118	.607

Table 2: Within-trait overlap between generative-AI and linguistic features.

4.2 RQ2: Overall Model Performance and Incremental Gains

Across the nine random forest models, combined features produced the highest QWK, exact agreement, and R^2 and the lowest MAE for all traits (Table 3), although absolute SMD was slightly larger than for linguistic-only models for Organization/Purpose and Evidence/Elaboration, and adjacent agreement was slightly lower for Conventions. Generative-AI-only models trailed linguistic-only models by .126–.132 of QWK and 6.5–11.1 percentage points in exact agreement. These deficits were substantially larger than the modest combined-model QWK gains of .011–.019, supporting the use of generative-AI features as supplemented predictors. Applying the benchmarks described by Williamson et al. (2012), all models met $|SMD| < .15$, but all Conventions models and

all generative-AI-only models fell below QWK .70; only linguistic-only and combined models were within .10 of human–human QWK. Combined-model and human–human QWK values were of similar magnitude, but they are not directly comparable because the models were fitted to Rater 1 scores, whereas human–human agreement was estimated on a smaller, partially overlapping sample of double-scored test essays.

Paired bootstrap intervals provided the only QWK interval excluding zero, marginally, for Organization/Purpose, with a gain of .019, 95% CI [.0010, .0376], as shown in Table 4. Intervals are conditional on one split and one fit, exclude model-fitting and feature-extraction variance, and are unadjusted across nine comparisons. Evidence/Elaboration and Conventions had positive estimates but intervals including zero. These outcomes do not establish statistically different augmentation effects between traits.

For Conventions, exact agreement increased by 1.97 percentage points and MAE decreased, with both intervals excluding zero. Neither secondary interval excluded zero for the other traits. These comparisons suggest exploratory, metric-specific benefits.

4.3 RQ2: Model Performance within Human Score Categories

Scoring category-level agreement showed where augmentation contributed to these gains (Table 5). For Organization/Purpose, exact agreement rose 3.0 percentage points at score 1 but fell at score 4, from 21.4% to 17.9% (6 to 5 of 28 essays). Evidence/Elaboration improved at scores 1, 3, and 4 but declined at score 2; the score-4 gain (22.2% to 25.9%) reflects 6 to 7 of 27 essays. Conventions gained most in the middle category, with a smaller upper-category improvement and unchanged lowest-category agreement. Because the

Trait	Features	p	QWK	Exact %	Exact + Adjacent %	SMD	MAE	R^2
Organization/ Purpose	Linguistic only	19	.734	72.01	99.49	-.008	.348	.618
	Generative-AI only	13	.608	63.63	98.62	-.022	.426	.461
	Combined	32	.753	73.47	99.71	-.030	.338	.640
	Human-human	—	.727	68.57	100.00	-.038	—	—
Evidence/ Elaboration	Linguistic only	32	.738	72.96	99.56	.003	.337	.635
	Generative-AI only	14	.608	61.88	98.47	-.090	.435	.472
	Combined	46	.749	73.54	99.64	-.023	.330	.646
	Human-human	—	.754	70.86	100.00	-.015	—	—
Conventions	Linguistic only	9	.658	69.61	99.56	.054	.374	.499
	Generative-AI only	15	.526	63.12	98.62	.052	.421	.361
	Combined	24	.673	71.57	99.42	.047	.363	.527
	Human-human	—	.632	66.57	100.00	-.029	—	—

Table 3: Performance evaluation of the nine random forest scoring models.

Note. Model metrics use $N = 1,372$ test essays; human-human metrics use $N = 350$ double-scored test essays. p = number of predictors; QWK = quadratic weighted kappa; Exact + Adjacent = agreement within one score point; SMD = standardized mean difference; MAE = mean absolute error using raw predictions. R^2 also uses raw predictions.

score-4 categories contain 28 and 27 essays, each percentage point there corresponds to well under one essay, and those rows describe single-essay shifts rather than reliable differences.

Mean predicted scores for the linguistic-only and combined models were similar across human-score categories for all three traits. For Organization/Purpose and Evidence/Elaboration, the combined model slightly reduced lowest-category overprediction, but highest-category underprediction persisted and was more pronounced for generative-AI-features-only models. The second human ratings also showed higher mean scores in the lowest reference-score category and lower mean scores in the highest category, indicating that this pattern was not unique to automated scoring. Overall, adding generative-AI features to the linguistic features produced limited changes in category mean predictions. Supplementary Figure S1 presents the confusion matrices for more details.

5 Discussion

Our findings support rubric-aligned generative-AI features as promising supplements to established linguistic features. Their prominent correlations with human scores for Organization/Purpose and Evidence/Elaboration provide preliminary evidence of relevance to human judgments. Partial overlap between the two feature sources suggests opportunities for combining their information. Although some highly correlated features may share information, this does not diminish their potential contribution; model comparisons establish whether their integration improves scoring prediction.

The combined models achieved higher QWK than linguistic-only models across all three traits, with gains of .011–.019. Generative-AI-only models performed less well, reinforcing the value of the proposed hybrid approach: retaining established linguistic features while incorporating additional rubric-aligned generative-AI features. These findings extend Eltanbouly et al.’s (2025) rubric-derived feature approach and align with earlier evidence supporting linguistic features in hybrid scor-

Trait	Metric	Difference	95% Confidence Interval
Organization/ Purpose	QWK	0.0190	[0.0010, 0.0376]
	Exact agreement (pp)	1.46	[−0.29, 3.21]
	MAE (rounded scores)	−0.0168	[−0.0343, 0.0007]
Evidence/ Elaboration	QWK	0.0110	[−0.0064, 0.0287]
	Exact agreement (pp)	0.58	[−0.95, 2.11]
	MAE (rounded scores)	−0.0066	[−0.0219, 0.0095]
Conventions	QWK	0.0157	[−0.0016, 0.0332]
	Exact agreement (pp)	1.97	[0.51, 3.43]
	MAE (rounded scores)	−0.0182	[−0.0328, −0.0036]

Table 4: Paired bootstrap differences between combined and linguistic-only models.

Note. QWK = quadratic weighted kappa; pp = percentage points; MAE = mean absolute error using rounded predictions. Differences are combined minus linguistic-only.

ing systems (Uto et al., 2020; Hou et al., 2025).

The clearest improvements occurred for Organization/Purpose QWK and Conventions exact agreement and rounded-score MAE. These results illustrate two useful patterns: Generative-AI features ranked prominently for Organization/Purpose, while their integration also benefited Conventions, where linguistic features dominated the strongest individual correlations. Features need not rank highest individually to contribute within a combined model. Evidence/Elaboration estimates also favored integration, although confidence intervals included zero. This pattern supports further refinement and replication without establishing a significant gain for that trait. Huang et al. (2026), who prompted generative-AI to assign trait scores directly, similarly reported different strengths across scoring approaches. Our findings support examining integration by trait and metric, while not establishing statistically different augmentation effects between traits.

At the score-category level, combined models slightly reduced lowest-category overprediction for Organization/Purpose and Evidence/Elaboration. Upper-category underprediction remained, identifying a target for development. Similar movement toward central scores in second human ratings suggests that this challenge extends beyond automated scoring. Broader representation of high-scoring essays could help determine whether our generative-AI features augmentation extend benefits further across the score range.

5.1 Limitations and Future Directions

Several limitations warrant consideration. First, this study examined only one writing prompt, and human–human agreement was derived from a smaller double-scored subset for descriptive context only. Furthermore, shrinkage toward central score categories and sparse representation at the highest score bands limited predictive precision. Future studies should therefore examine multiple prompts with broader and more balanced score distributions. Additionally, the lack of demographic data precluded subgroup fairness analyses, highlighting a critical area for future evaluation.

Second, the generative-AI feature-based scoring process omitted source passages and confidential prompt materials. Consequently, Evidence/Elaboration features could not verify source accuracy or completeness, rendering them potentially vulnerable to surface-level citation coaching. Future research could incorporate prompt and passage context to ensure generative-AI feature scores capture authentic content integration.

Third, methodological stability remains unverified. Relying on a single codebook generation and model run left generative-AI features’ reliability untested. Subsequent work should assess codebook consistency across repeated generations and evaluate feature scoring stability. Crucially, studies controlling for predictor counts and model configurations are needed to clarify whether gains stem uniquely from generative-AI features rather than added predictor capacity. It’s important to note that

Trait	Human Score	<i>N</i>	Generative-AI only (%)	Linguistic only (%)	Combined (%)	Change (pp)
Organization/Purpose	1	603	68.8	78.1	81.1	3.0
	2	533	71.3	75.2	75.8	0.6
	3	208	37.5	52.9	52.9	0.0
	4	28	0.0	21.4	17.9	−3.5
Evidence/Elaboration	1	600	74.0	77.7	80.5	2.8
	2	542	58.7	77.3	74.7	−2.6
	3	203	42.9	54.2	56.2	2.0
	4	27	0.0	22.2	25.9	3.7
Conventions	0	180	25.6	45.0	45.0	0.0
	1	427	61.8	60.7	65.1	4.4
	2	765	72.7	80.4	81.4	1.0

Table 5: Exact agreement by human score category across models.

Note. *N* = number of test essays; pp = percentage points. Change represents combined minus linguistic-only agreement. Score-4 rows (*n* = 28 and 27) are sensitive to single essays: one essay corresponds to approximately 3.6 and 3.7 percentage points, respectively.

these generative-AI features and models are only research prototypes and are not intended for operational scoring decisions without further validation.

Finally, further systematic content analysis is needed to examine how linguistic and generative-AI features conceptually align. Investigating where these representations overlap or complement one another will clarify their capacity to support interpretable scoring and inform accurate, actionable automated feedback for student revision.

References

- John Atkinson and Diego Palma. 2025. [An LLM-based hybrid approach for enhanced automated essay scoring](#). *Scientific Reports*, 15:14551.
- Yigal Attali and Jill Burstein. 2006. [Automated essay scoring with e-rater V.2](#). *The Journal of Technology, Learning and Assessment*, 4(3).
- Leo Breiman. 2001. [Random forests](#). *Machine Learning*, 45:5–32.
- Michael B. Bunch, Thomas Davis, Ann Hayes, Derek Justice, and Julie St. John. 2017. White paper: PEG changes. White paper, Measurement Incorporated.
- SeongYeub Chu, Jong Woo Kim, Bryan Wong, and Mun Yong Yi. 2025. [Rationale behind essay scores: Enhancing S-LLM’s multi-trait essay scoring with rationale generated by LLMs](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5811–5829. Association for Computational Linguistics.
- Heejin Do, Sangwon Ryu, and Gary Geunbae Lee. 2025. [Teach-to-reason with scoring: Self-explainable rationale-driven multi-trait essay scoring](#). *arXiv Preprint*, arXiv:2502.20748.
- Sohaila Eltanbouly, Salam Albatarni, and Tamer Elsayed. 2025. [TRATES: Trait-specific rubric-assisted cross-prompt essay scoring](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20528–20543. Association for Computational Linguistics.
- Zhaoyi Hou, Alejandro Ciuba, and Xiang Li. 2025. [Improving LLM-based automatic essay scoring with linguistic features](#). In *Proceedings of the Innovation and Responsibility in AI-Supported Education Workshop*, volume 273 of *Proceedings of Machine Learning Research*, pages 41–65.
- Yue Huang, Corey Palermo, Ruitao Liu, and Yong He. 2025. [An early review of generative language models in automated writing evaluation: Advancements, challenges, and future directions for automated essay scoring and feedback generation](#). *Chinese/English Journal of Educational Measurement and Evaluation*, 6(2):5.
- Yue Huang, Corey Palermo, and Joshua Wilson. 2026. [Accuracy and fairness of generative AI in automated essay scoring: Comparing GPT-4o, feature-based models, and human raters](#). *Assessing Writing*, 69:101047.
- Yue Huang and Joshua Wilson. 2025. [Evaluating LLM-based automated essay scoring: Accuracy, fairness, and validity](#). In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Works in Progress*, pages 71–83, Pitts-

burgh, Pennsylvania, United States. National Council on Measurement in Education (NCME).

- Syed Latifi and Mark Gierl. 2021. [Automated scoring of junior and senior high essays using Coh-Metrix features: Implications for large-scale language testing](#). *Language Testing*, 38(1):62–85.
- Danielle S. McNamara, Scott A. Crossley, Rod D. Roscoe, Laura K. Allen, and Jianmin Dai. 2015. [A hierarchical classification approach to automated essay scoring](#). *Assessing Writing*, 23:35–59.
- Ellis B. Page. 1966. The imminence of...grading essays by computer. *Phi Delta Kappan*, 47(5):238–243.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. [Scikit-learn: Machine learning in Python](#). *Journal of Machine Learning Research*, 12:2825–2830.
- Mark D. Shermis. 2014. [State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration](#). *Assessing Writing*, 20:53–76.
- Smarter Balanced Assessment Consortium. n.d. [Argumentative performance task writing rubric \(Grades 6–11\)](#).
- Tamara P. Tate, Jacob Steiss, Drew Bailey, Steve Graham, Youngsun Moon, Daniel Ritchie, Waverly Tseng, and Mark Warschauer. 2024. [Can AI provide useful holistic essay scoring?](#) *Computers and Education: Artificial Intelligence*, 7:100255.
- Masaki Uto, Yikuan Xie, and Maomi Ueno. 2020. [Neural automated essay scoring incorporating hand-crafted features](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6077–6088. International Committee on Computational Linguistics.
- David M. Williamson, Xiaoming Xi, and F. Jay Breyer. 2012. [A framework for evaluation and use of automated scoring](#). *Educational Measurement: Issues and Practice*, 31(1):2–13.

A Supplementary Figures

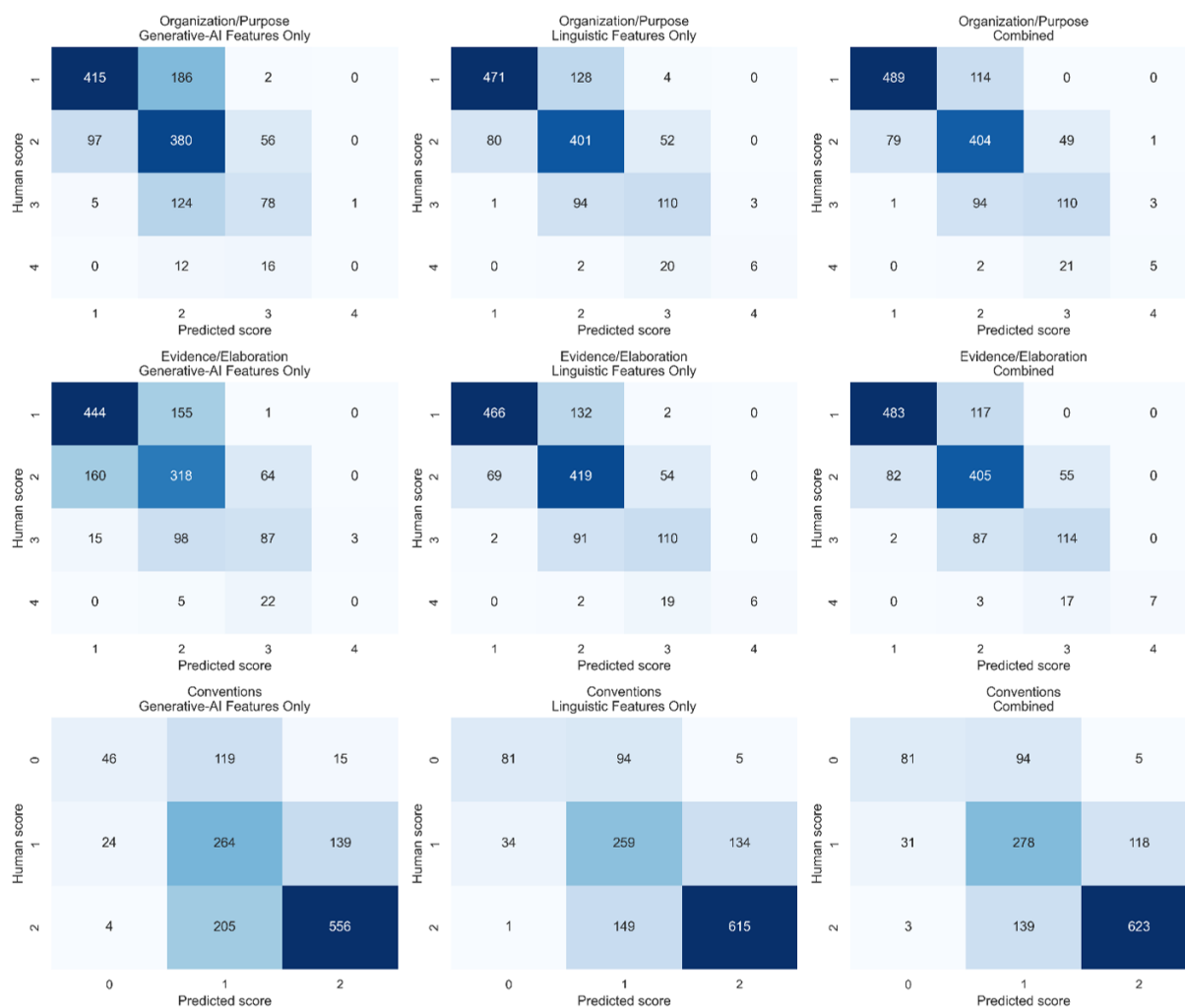


Figure S1: Confusion matrices for random forest trait scores across the three feature conditions