Explaining Holistic Essay Scores in Comparative Judgment Assessments by Predicting Scores on Rubrics

Michiel De Vrindt [1ac] **Anaïs Tack** [1ad] **Renske Bouwer** [2b]
Wim Van den Noortgate [1ac] **Marije Lesterhuis** [3e]

[1] KU Leuven [a] imec research group itec [b] Institute for Language Sciences
[c] Faculty of Psychology and Educational Sciences [2] Utrecht University [3] UMC Utrecht
[d] Faculty of Arts [e] Center for Research and Development of Health Professions Education

Abstract

Comparative judgment (CJ) is an assessment method in which multiple assessors determine the holistic quality of essays through pairwise comparisons. While CJ is recognized for generating reliable and valid scores, it falls short in providing transparency about the specific quality aspects these holistic scores represent. Our study addresses this limitation by predicting scores on a set of rubrics that measure text quality, thereby explaining the holistic scores derived from CJ. We developed feature-based machine learning models that leveraged complexity and genre features extracted from a collection of Dutch essays. We evaluated the predictability of rubric scores for text quality based on linguistic features. Subsequently, we evaluated the validity of the predicted rubric scores by examining their ability to explain the holistic scores derived from CJ. Our findings indicate that feature-based prediction models can predict relevant rubric scores moderately well. Furthermore, the predictions can be used to explain holistic scores from CJ, despite certain biases. This automated approach to explain holistic quality scores from CJ can enhance the transparency of CJ assessments and simplify the evaluation of their validity.

1 Introduction

Comparative judgment (CJ) is a widely used method for educational assessments, particularly for evaluating writing quality of essays (Baniya et al., 2019; Steedle and Ferrara, 2016; van Daal et al., 2016). In CJ, assessors repeatedly compare (different) pairs of essays and determine which one is superior in quality each time. Then, the Bradley-Terry-Luce (BTL) model (Bradley and Terry, 1952; Luce, 1959), relates the probability of one essay being preferred over another to the quality scores of the essays that are compared. Based on the judgments of assessors, the quality scores of essays are estimated.

CJ offers several advantages over traditional rubric-based assessments. Firstly, it allows assessors to use their professional expertise and intuition without strictly adhering to predetermined rubrics, making CJ a more natural assessment method (Bloxham, 2009; Laming, 2003). Assessors may have different conceptualizations of quality; some prioritize essay argumentation and organization, while others focus on language conventions (Lesterhuis et al., 2022). Even when assessors focus on different aspects, van Daal et al. (2016) found that their pairwise comparisons still reflected construct-relevant aspects of writing quality. Secondly, since CJ incorporates multiple judgments from various assessors, the resulting essay quality scores are generally reliable and valid, reflecting a consensus among the assessors (Lesterhuis et al., 2022; Verhavert et al., 2019; van Daal et al., 2016). Although CJ is a valid and reliable assessment method, the holistic scores it produces lack transparency regarding their specific meaning. Since judgments are made holistically, the assessors' decision-making process remains unclear. Assessors can provide feedback while making judgments, but this takes more time and may shift their focus from the overall quality of essays to specific analytic criteria (Verhavert et al., 2019). Finding a new way to explain holistic scores is therefore crucial for making CJ assessments more transparent and can also serve as a form of feedback.

In this study, we investigate the use of feature-based prediction models to explain holistic quality scores from CJ, with the goal of enhancing their transparency.

Our research addresses the following questions:

1. How reliably can scores on a set of rubrics measuring text quality be predicted based on linguistic features of essay texts?

2. To what extent do these predicted rubric scores accurately reflect the holistic quality

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 535–548
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

scores of essays obtained with CJ?

Our study comprised two phases. First, we conducted a machine learning experiment to assess how well rubric scores could be predicted from linguistic features of Dutch essays. Second, we performed a regression analysis to evaluate the validity of the predicted scores in explaining the holistic scores obtained with CJ.

2 Background

2.1 Comparative Judgment Assessments

CJ functions as an alternative assessment method to rubric scoring and has been shown to produce reliable and valid scores (Verhavert et al., 2019; Lesterhuis et al., 2022; van Daal et al., 2016; Heldsinger and Humphry, 2010). While primarily known for assessing essay quality, CJ has also been effectively used for various other types of assessments. These include evaluating conceptual understanding (Jones et al., 2019), mathematical problem-solving skills (Jones and Inglis, 2015), design portfolios (Newhouse, 2014), formative assessments (Potter et al., 2017; Bartholomew et al., 2019), and comparing assessment standards across examination boards (Bramley, 2007; D'Arcy, 1997).

Generally, CJ assessments are conducted by iterating through three key steps. In the first step, a pair of essays is chosen and assigned to one of several assessors. In the second step, the assessor compares the two essays and determines which demonstrates higher quality. This relative assessment approach is considered more intuitive than absolute assessments, such as rubric-based scoring. As Laming (2003) noted, all judgments inherently involve comparing one entity to another, and CJ explicitly makes use of this principle. In the third step, the BTL model is applied to link the outcomes of all pairwise comparisons to a quality scale (Bradley and Terry, 1952; Luce, 1959). The BTL model relates the probability of one essay being favored over another to the difference in their quality scores, expressed as logit values. Specifically, this probability is determined by the sigmoid function of the quality score difference: the greater the quality score of the first essay relative to the second, the higher the probability it will win the comparison. The quality scores in BTL model are continuously updated based on the judgments that assessors make. The assessment concludes once a sufficient number of judgments have been collected, typically requiring each essay to be compared 10 to 14 times to ensure reliable quality scores. Ultimately, the holistic scores derived from CJ are both reliable and valid, as they stem from numerous pairwise comparisons by multiple assessors (van Daal et al., 2016; Lesterhuis et al., 2022).

However, when the CJ assessment is completed, the resulting quality scores for essays lack clarity regarding what they represent. The issue stems from the scores being based on holistic pairwise comparisons by assessors (Steedle and Ferrara, 2016; Kelly et al., 2022), a method that, while reliable and valid, lacks the transparency offered by detailed rubric-based marking (Jonsson, 2014; Mortier et al., 2015). As a result of this ambiguity, the feedback function of the scores to students is hindered, and the validation of the assessors' judgments is complicated. Even though assessors can provide feedback comments when making judgments, doing so extensively would be time-consuming and reduce assessment efficiency. Furthermore, writing numerous comments to individual essays can lead assessors to adopt a more analytical approach (Verhavert et al., 2019), which conflicts with the holistic nature of CJ assessments (van Daal et al., 2016). Hence, there is a need to enhance the transparency of the holistic scores obtained with CJ without requiring more effort from assessors. To achieve this, we propose automatically predicting the scores on rubrics to explain the holistic scores derived from CJ. This prediction task is similar to that of automated essay scoring (AES).

2.2 Automated Essay Scoring

With recent advancements in NLP methods, AES for summative assessments and automatic writing evaluation (AWE) for formative assessments have received increasing attention. Initially, systems relied on analyzing hand-crafted linguistic features from essay texts to predict scores (Ke and Ng, 2019). However, following the ASAP Kaggle competition organized by the Hewlett Foundation (Hamner et al., 2012), deep learning models have gained prominence in this domain, often surpassing traditional feature-based prediction models in terms of agreement with human scoring (Dong et al., 2017; Taghipour and Ng, 2016; Wang et al., 2022). Despite these advances, practical AES and AWE systems, such as PEG (Dikli, 2006) and e-rater (Burstein et al., 2004), continue to rely heavily on hand-crafted linguistic features due to the need for transparency. Especially, text complexity features such as syntactical complexity and lexical diversity have been shown to have a large predictive power for the writing quality of essays (McNamara et al., 2010). For instance, for English-written essays, the Coh-Matrix (Graesser et al., 2004) and SALAT toolsets (Crossley et al., 2023) are commonly used to extract complex linguistic features for AES (McNamara et al., 2015; Li and Liu, 2017; Latifi and Gierl, 2021; Kumar and Boulanger, 2020).

While feature-based AES models and AWE systems provide more transparency, the linguistic features themselves, such as complexity and cohesion features, can still be hard to interpret and may lack pedagogical clarity for students and teachers. As Deane (2013b) stated, using linguistic features as proxies for writing quality is neither transparent nor instructional for students. Additionally, Crossley (2020) noted that extensive knowledge is required in order to use linguistic features effectively.

As assessors mostly consider higher-order aspects of writing when making pairwise comparisons, such as structure and argumentation (Lesterhuis et al., 2022), linguistic features would not provide the desired transparency about the holistic quality scores. Therefore, in this study, we chose to explain the holistic scores based on more instructional rubrics that measure specific aspects of writing quality. We predicted scores across these rubrics based on linguistic features extracted from essays. The automated scoring task of predicting scores across multiple rubrics is also referred to as 'multi-trait' scoring within AES literature (He et al., 2022; Do et al., 2023; Mathias and Bhattacharyya, 2020).

3 Method

3.1 Data

We used data previously collected by Coertjens et al. (2017). The dataset, detailed in Table 1, included a total of 104 argumentative essays in Dutch written by students from secondary education. The students could choose to write an essay on one of the topics: (1) having children, (2) organ donation, and (3) stress experienced by students. Despite the differences in topics, the essays were quite similar in terms of the assessed competence: the ability to effectively integrate source material within argumentative writing. This allowed us to combine the essays from different assignments into one dataset for model training. We selected this data because it is the only CJ dataset where essays are labeled with both holistic and rubric scores.[1]

Assignment	Essays	Tokens	Tokens/Essay
	N		$M \pm SD$
1. Children	34	11167	328 (± 92)
2. Organ	35	11358	293 (± 93)
3. Stress	35	11859	304 (± 97)

Table 1: Overview of the argumentative writing assignment gathered by Coertjens et al. (2017). Tokenization was performed using the Dutch `nl_core_news_sm` model from spaCy (Explosion, 2023).

3.1.1 Holistic Scores

Coertjens et al. (2017) used CJ to obtain holistic scores of essay quality. During the assessment, 40 assessors made pairwise comparisons and each essay was compared 25 times. The assessors were asked which essay in this pair is better in terms of argumentation. This assessment resulted in holistic scores with a reliability of 0.87, as measured by scale separation reliability (Verhavert et al., 2018).

3.1.2 Scores on Rubrics

Coertjens et al. (2017) asked 18 assessors to evaluate the same essays using a rubric set designed to measure 20 aspects of text quality. These were different assessors from those who scored the essays holistically with CJ. These aspects were grouped into four main components: structure (6 rubrics), content (7 rubrics), argumentation (4 rubrics), and language conventions (3 rubrics). The rubrics, originally developed and validated by Rijlaarsdam et al. (1994), were adapted by Coertjens et al. (2017) for this particular assignment on argumentative writing. According to Coertjens et al. (2017), the intraclass correlation coefficient was 0.85 after five different assessors assessed each essay. For an overview and description of all rubrics, refer to Appendix A.

3.2 Features

To extract linguistic features from the essays, we used T-Scan (Maat et al., 2014) because Dascalu et al. (2017) previously demonstrated that its features have strong predictive power for automated essay scoring. Using the T-Scan API (v0.10), we extracted 476 document-level features related to lexical complexity, sentence complexity, referential cohesion, lexical diversity, lexical semantics,

[1] The data gathered by Lesterhuis et al. (2022), for example, includes a superset of the essays used by Coertjens et al. (2017), but it does not include any rubric scores.

and personal style. For details on the T-Scan configuration, see Appendix B.

Since T-Scan does not account for spelling and grammatical errors, we also used the LanguageTool package (v2.8.1) in Python to count the number of language mistakes in each essay. We normalized these counts by dividing them by the total number of tokens per essay (see Table 1).

3.3 Models

We trained regression models using the extracted features to predict scores of essay quality. This involved training multiple single-target regression models, with each model predicting either the holistic score or one of the rubric scores.

We experimented with five machine learning models for the regression tasks: **Lasso Regression**, **ElasticNet**, **Random Forest**, and **XGBoost** using scikit-learn 1.4.0 (Pedregosa et al., 2011), and **LightGBM** 4.6.0 (Ke et al., 2017) in Python 3.9.12. We applied min-max normalization to each input feature from T-Scan as well as the rubric scores. Before training the models, we excluded features from the training set that had a low Pearson correlation with the target. A correlation threshold of 0.12 was chosen based on Lovakov and Agadullina (2021).

For each individual model, we ran hyperparameter tuning on the training set using a randomized search strategy with up to 100 iterations. The optimal hyperparameters were selected based on the lowest mean absolute error (MAE) between the predicted and actual rubric scores across 20 folds.

3.4 Evaluation

To optimize the prediction performance and avoid overfitting on a small dataset, we performed leave-one-out cross-validation (LOOCV). This involved leaving out one essay for evaluation and training a model on all remaining essays, repeating the process for each essay in the dataset. The hyperparameter tuning with 20-fold cross-validation, as mentioned before, was conducted on the training data for each run of LOOCV. Refer to Appendix C for an overview of the selected hyperparameters.

3.4.1 Metrics

Using the optimal model and hyperparameters for each rubric, we evaluated the predictions of the rubric scores with various metrics on all left-out essays during LOOCV. We used the **squared Pearson correlation coefficient** (R^2) between predicted and actual scores. R^2 is a measure of score reliability in classical test theory (Brennan, 2010) and is often used to measure the reliability of quality scores estimated from CJ relative to true scores (Verhavert et al., 2018).

Additionally, we used the **quadratic weighted kappa** (QWK) (Cohen, 1968) and the **mean absolute error** (MAE), two commonly used metrics in AES research (Ramesh and Sanampudi, 2022). QWK is a metric based on Cohen's kappa that measures agreement between predicted and human-given scores, penalizing more divergent predictions. A score of 1 indicates perfect agreement, while -1 indicates perfect disagreement.

3.4.2 Predictive Power

To validate whether the predicted rubric scores accurately measure the assessed writing quality with CJ, we evaluated their predictive power for the holistic scores using linear regression models. Using the statsmodels package (0.13.2) (Seabold and Perktold, 2010), we constructed two regression models measuring the effects of rubric scores on holistic scores from CJ:

- **Regression Model 1** uses the rubric scores predicted by the model (see Section 3.3) as covariates and holistic quality scores from CJ as outcomes.
- **Regression Model 2** uses the rubric scores given by assessors as covariates and holistic quality scores from CJ as outcomes.

We compared their goodness-of-fit using the Akaike information criterion (AIC) and Bayesian information criterion (BIC), as well as the explained variance (R^2) of the holistic scores. In the context of statistical modeling, R^2 measures the proportion of variance in holistic scores that is explained by the (predicted) scores on rubrics. As it can be inflated by adding many covariates, we adjusted R^2 for the number of covariates.

We expected that Model 2, which uses human-assigned rubric scores, would fit the holistic scores better than Model 1, which uses predicted rubric scores. Therefore, we aimed to evaluate how closely the fit of Model 1 approximates that of Model 2.

To further investigate any potential biases in the effects (i.e., coefficients) of the predicted rubric scores on the holistic scores, we compared the coefficients of Model 1 with those of Model 2, along with their confidence intervals. We calculated Student's t-tests to evaluate whether the coefficients are significantly different from zero. We omitted intercepts for both models to make the coefficients comparable, and inverted the normalization of all variables to their original scales.

4 Results

4.1 Predicting Scores on Rubrics

Table 2 shows the performance of the rubric scoring models evaluated using LOOCV. ElasticNet consistently demonstrated the best performance in predicting most rubric scores (9/20 rubrics), followed by XGBoost (5/20 rubrics), Lasso (3/20 rubrics), RandomForest (2/20 rubrics), and LightGBM (1/20 rubrics). Overall, model performance was moderately effective across all evaluation metrics, which is expected given the small sample size.

Performance varied notably across the different rubrics. Among all rubrics pertaining to essay structure, predictions showed the highest reliability (R^2) for Construction, Relationships, and Continuity, and the highest agreement (QWK) with human scores for Title compared to other rubrics. This suggests that the linguistic features employed in the models can capture markers of essay organization, despite the overall moderate prediction performance.

Among rubrics pertaining to essay content, scores on References and Citations were the most accurately predicted, but generally, the predictions were only moderate or poor. The scores for Introduction, Persuasion, Reader Focus, Reader Engagement, and Conclusion were comparatively worse. This suggests that the linguistic features employed in the models can capture some markers of essay content, albeit with limited accuracy.

Among rubrics pertaining to argumentation, Support and Relevance showed the best prediction performance, whereas Indication and Reference Cohesion Relationships were less accurately predicted. This shows the model's ability to capture, to a certain extent, the argumentative writing level related to how sources were integrated and used to support claims.

Generally, the rubrics pertaining to language were poorly predicted. Among these rubrics, the scores for Style were most accurately predicted, while predictions for Grammar and Spelling, and Punctuation were comparatively worse. Hence, the assessors' scoring of language conventions differed from the model predictions.

4.2 Explaining Holistic Scores with Predictions

After evaluating the predicted scores on rubrics, we examined how well these scores can explain the holistic scores obtained through CJ. Table 3 shows the goodness-of-fit of the two models. As expected, Model 2 provided a better fit for the holistic scores than Model 1, as evidenced by smaller AIC and BIC values. Additionally, Model 2 explained 12% more of the variance in holistic scores compared to Model 1, as indicated by the higher R^2. This difference was even more pronounced when considering the adjusted R^2 values. Although the predicted rubric scores explained the holistic scores reasonably well (Model 1), 40% of the variance in holistic scores still remained unaccounted for. In contrast, the rubric scores given by assessors (Model 2) had greater predictive power for the holistic scores. However, even in Model 2, 28% of the variance in holistic scores remained unexplained, indicating a difference in how assessors score essays with rubrics versus holistically using CJ.

Figure 1 illustrates the biases in the coefficients of the predicted rubrics on the holistic scores (Model 1) with respect to the coefficients of the assessor-assigned rubric scores (Model 2). Overall, the coefficients for the rubrics in both models were similar in magnitude and direction, which supports the validity of the predicted rubric scores. However, Model 1 exhibited some systematic biases, as it tends to overshoot the magnitude of the coefficients. Specifically, the coefficients for rubrics pertaining to essay structure showed upward biases, except for Subtopic. Conversely, the coefficients for rubrics related to content displayed downward biases, with the exception of Introduction and Citations.

The most significant coefficients in Model 2 were Relationships, References, Conclusion, and Grammar and Spelling. Except for Grammar and Spelling, these rubrics were all reasonably well approximated by Model 1, as their coefficients were similar and their confidence intervals overlapped. This shows that predicted rubric scores accurately explained the holistic scores based on the most important rubrics scored by assessors. However, of these rubrics, only the coefficients for Relationships, and Grammar and Spelling were significantly different from zero in Model 1. This can be attributed to the wider confidence intervals of Model

Aspect	Rubric	Best Model	R^2	QWK	MAE
Structure	Title	LightGBM	0.23	0.50	0.85
Structure	Construction	XGBoost	0.31	0.41	0.80
Structure	Layout	RandomForest	0.24	0.41	0.91
Structure	Subtopic	XGBoost	0.25	0.41	0.74
Structure	Relationships	ElasticNet	0.30	0.46	0.76
Structure	Continuity	RandomForest	0.34	0.45	0.48
Content	Introduction	ElasticNet	0.27	0.45	0.54
Content	Persuasion	ElasticNet	0.31	0.45	0.53
Content	References	Lasso	0.48	0.60	0.58
Content	Citations	XGBoost	0.53	0.64	0.54
Content	Reader Focus	ElasticNet	0.30	0.40	0.39
Content	Reader Engagement	ElasticNet	0.38	0.42	0.38
Content	Conclusion	XGBoost	0.28	0.46	0.66
Argumentation	Support	Lasso	0.51	0.59	0.35
Argumentation	Relevance	ElasticNet	0.46	0.54	0.36
Argumentation	Indication	ElasticNet	0.22	0.30	0.59
Argumentation	Reference Cohesion Relationships	XGBoost	0.28	0.37	0.47
Language	Grammar and Spelling	Lasso	0.25	0.43	0.48
Language	Punctuation	ElasticNet	0.27	0.35	0.49
Language	Style	ElasticNet	0.38	0.45	0.34

Table 2: Evaluation of LOOCV results for predicting scores on rubrics measuring aspects of text quality, with hyperparameter search conducted for each run.

Reg. Model	AIC	BIC	R^2	Adj. R^2
1	427.60	483.20	0.60	0.50
2	390.20	445.80	0.72	0.65

Table 3: Comparison of model fit for linear regression models with holistic scores from CJ as the outcome, where Regression Model 1 used predicted scores on rubrics as covariates and Regression Model 2 used scores on rubrics provided by assessors as covariates.

1 compared to Model 2.

When examining rubrics that were most accurately predicted (see Table 2), it is clear that their impact on holistic scores (Model 1) closely resembled that of scores given by assessors (Model 2). This similarity was evident for Relationships, References, and Citations, where Model 1's coefficients aligned with those of Model 2, and their confidence intervals significantly overlapped. Although Support and Relevance were also predicted more accurately, their coefficients exhibited great uncertainty, as indicated by wider confidence intervals compared to Model 2, especially for Support. This indicates a potential lack of validity of the predicted rubrics related to argumentation.

Conversely, when the rubric was predicted more inaccurately, their coefficients showed more bias. This was observed for Layout and Reader Focus, as their rubric scores were poorly predicted and their coefficients overestimated. Similarly, the rubrics related to language were inaccurately predicted when compared to other rubrics, resulting in coefficients with large biases. More specifically, the importance of sound Grammar and Spelling was vastly overestimated using the predicted rubric scores, as these scores are much higher than when rubric scores were given by teachers. Conversely, the importance of correct Punctuation and Style was overly negative when using the predictions compared to the scores given by teachers. This shows that teachers score rubrics differently and that these scores contribute less to the holistic scores than when using predictions.

5 Discussion

The lack of transparency in holistic scores obtained through CJ limits the practical application of this assessment method (Steedle and Ferrara, 2016). Our analysis reveals that rubric scores can be predicted moderately well in terms of reliability and

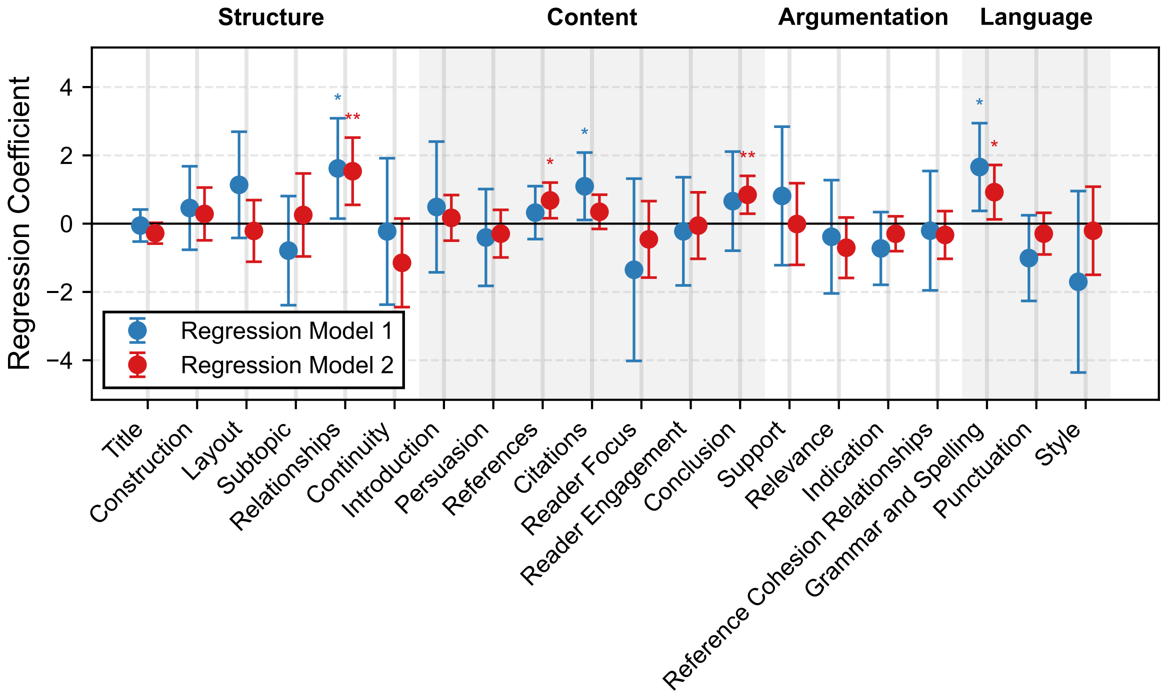

Figure 1: Comparison of regression coefficients of Model 1 and Model 2 for all rubric scores with 95% confidence intervals, where significance is denoted as * $p < 0.05$, ** at $p < 0.01$, and *** at $p < 0.001$.

agreement with assessor-assigned scores, with the best predictions for the rubrics relevant to the assignment on argumentative writing. However, not all rubrics can be predicted well, which can be due to the limited size of the training set.

Furthermore, the predicted rubric scores have explanatory power for holistic scores derived from CJ, thus showing potential for an automated approach to provide more transparency. However, it is unlikely that the rubric scores can fully explain the differences in holistic scores from CJ, as even the assessor-assigned scores do not fully explain them. This is not surprising, given that holistic scoring with CJ involves relative assessments while rubric-based scoring requires absolute assessments. While both assessment approaches are reliable on their own, they may yield slightly different results (Coertjens et al., 2017).

Generally, the relationship between predicted rubrics and holistic scores is similar to that of rubric scores given by assessors, which supports the validity of the predictions. Most importantly, we find that the validity of the predicted rubric scores depends on their predictability from linguistic features. The rubrics that demonstrate better predictability can also explain the holistic scores with minimal bias. This shows that predicted scores on assignment-relevant rubrics can explain, in part, the holistic scores from CJ.

However, there are notable differences between human scoring and automated scoring (Ben-Simon and Bennett, 2007). Previously, Ramineni and Williamson (2018) found that the e-rater AES system often overvalues organization while undervaluing content. Our findings generally support this, as most structure-related rubrics exert an overly positive influence on holistic scores when predicted compared to when scored by assessors, whereas most content-related rubrics show an overly negative influence when predicted compared to when assessed by humans.

While certain argumentation-related rubrics can be predicted comparatively well, their influence on the holistic scores shows more uncertainty. As Attali (2007) stated, agreement between human and automated scores does not directly imply that the scores are valid. This discrepancy may be attributed to the inherent difficulty in measuring the quality of argumentation based on linguistic features (Deane, 2013a). Hence, more elaborate linguistic features based on argument mining may be needed for more valid predictions of argumentative-related rubrics.

Additionally, there is a difference between how language-related rubrics are predicted and how they are assessed. These rubrics prove difficult to predict based on linguistic features, and their effect on holistic quality scores is biased, either undervalued or overvalued. Previously, Ramineni and Williamson (2018) noted that the e-rater AES system severely undervalues grammatical mistakes for essay scoring. Further analysis is needed to uncover the potential causes of this bias. However, for CJ assessments, language conventions are generally less important when making pairwise comparisons of essays (Lesterhuis et al., 2022).

6 Conclusion

To address the lack of transparency of holistic scores from CJ assessments, we used feature-based models to predict scores on a set of rubrics that explain the holistic scores. Based on linguistic features extracted with T-Scan, rubric scores of Dutch essays were predicted with moderate success. However, we found that the most relevant rubrics were predicted more reliably compared to other rubrics. Furthermore, we noted that these predicted scores on rubrics can explain holistic scores from CJ in a manner comparable to the assessor-assigned rubric scores.

While the automated predictions of rubrics offer more transparency regarding the meaning of holistic scores, they do differ from human assessor scores in certain respects. For instance, structure-related rubrics were slightly overvalued, content-related rubrics were slightly undervalued, and the effect of argumentation-related rubrics showed more uncertainty. Additionally, predictions for language convention rubrics diverged notably from assessor-given scores.

Despite some discrepancies in how predicted rubric scores explain holistic scores compared to rubrics scored by assessors, they generally aligned well for the most important rubrics and demonstrate predictive power. This suggests that predicting scores on rubrics can help explain the holistic scores obtained with CJ. However, their acceptance and effectiveness as feedback for students require future research.

Limitations

Even though the scores given by assessors are reliable and valid, the size of the available dataset used for training is rather limited, which could explain the moderate prediction performance. We expect that increasing the dataset would improve prediction performance and, therefore, produce scores on rubrics that better explain the holistic scores from CJ. With a larger training set, it would be possible to determine the best-performing hyperparameters and models for each rubric for all essays, rather than per fold as was done in this study. This approach would enhance the generalizability of the models' performance.

Future research could, for example, leverage the larger ASAP dataset, which contains English essays scored on rubrics such as ideas, organization, style, and conventions, for different writing genres (Hamner et al., 2012). However, the granularity of features is higher in this dataset, which would provide less specific explanations than in the current study.

In case only a small set of rubric-scored texts is available, it may be more suitable to extract rubric scores using language models, which can capture complex textual features. Large Language Models (LLMs) have been applied for this purpose through fine-tuning (Do et al., 2024) or zero-shot prompting (Lee et al., 2024). However, relying on LLMs would make it less transparent how the predicted scores are derived compared to using hand-crafted linguistic features, as in this study.

To better understand the validity of predicted scores on rubrics in relation to how they are predicted, future research could examine the most important features for making these predictions. This is important as feature-based approaches for AES do not capture meaning directly. Previously, it has been noted that essay length is highly influential for AES models, and it is advised that its effect be studied by controlling for it (Chodorow and Burstein, 2004). Text length is especially influential for structure, content, and argumentation, and less so for language (Enright and Quinlan, 2010; Barkaoui and Woodworth, 2023). While essay length is a valid factor that human assessors also consider, its disproportionate influence can be problematic. Moreover, analyzing the importance of linguistic features in the model's predictions can help clarify why language-related rubrics were predicted with low reliability and validity. This can be achieved by evaluating whether features extracted with LanguageTool significantly contributed to the prediction of language-related rubric scores.

Additionally, interpreting the coefficients of a linear regression model as effects of rubric scores on holistic scores requires caution. Rubrics measuring aspects of text quality tend to be highly correlated, which raises the potential for multicollinearity. To gain a more accurate understanding of how rubric scores influence the holistic scores, we recommend employing regularization techniques or incorporating interaction effects into the model. These approaches can help mitigate the challenges posed by correlated predictors and provide clearer insights into the effects of rubric scores on holistic scores from CJ. Furthermore, the validity of the predicted scores on rubrics for explaining holistic scores is contingent on the assessment context. In second language (L2) writing, for instance, criteria such as language accuracy may be weighted more heavily than they were in the L1 context of this study. Therefore, future research is essential to validate these findings across different assessment contexts.

Acknowledgments

This research was supported by a Baekeland mandate (HBC.2022.0164) granted by the Flanders Innovation & Entrepreneurship (VLAIO) to the first author, in cooperation with the company Comproved (D-Pac BV).

References

Yigal Attali. 2007. Construct validity of e-rater® in scoring toefl® essays. *ETS Research Report Series*, 2007(1):i–22.

Sweta Baniya, Nathan Mentzer, Scott R Bartholomew, Amelia Chesley, Cameron Moon, and Derek Sherman. 2019. Using adaptive comparative judgment in writing assessment. *The Journal of Technology Studies*, 45(1):24–35.

Khaled Barkaoui and Johanathan Woodworth. 2023. An exploratory study of the construct measured by automated writing scores across task types and test occasions. *Studies in Language Assessment*, 12(1):26.

Scott R Bartholomew, Greg J Strimel, and Emily Yoshikawa. 2019. Using adaptive comparative judgment for student formative feedback and learning during a middle school design project. *International Journal of Technology and Design Education*, 29:363–385.

Anat Ben-Simon and Randy Elliot Bennett. 2007. Toward more substantively meaningful automated essay scoring. *The Journal of Technology, Learning and Assessment*, 6(1).

Sue Bloxham. 2009. Marking and moderation in the UK: false assumptions and wasted resources. *Assessment & Evaluation in Higher Education*, 34(2):209–220.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Tom Bramley. 2007. Paired comparison methods. In Peter Newton, J Baird, H Goldstein, H Patrick, and P Tymms, editors, *Techniques for monitoring the comparability of examination standards*, pages 246–300. Qualifications and Curriculum Authority London, London, United Kingdom.

Robert L Brennan. 2010. Generalizability theory and classical test theory. *Applied Measurement in Education*, 24(1):1–21.

Jill Burstein, Martin Chodorow, and Claudia Leacock. 2004. Automated essay evaluation: The criterion online writing service. *AI Magazine*, 25(3):27–27.

Martin Chodorow and Jill Burstein. 2004. Beyond essay length: evaluating e-rater®'s performance on toefl® essays. *ETS Research Report Series*, 2004(1):i–38.

Liesje Coertjens, Marije Lesterhuis, San Verhavert, Roos Van Gasse, and Sven De Maeyer. 2017. Teksten beoordelen met criterialijsten of via paarsgewijze vergelijking: een afweging van betrouwbaarheid en tijdsinvestering. *Pedagogische studiën*, 94(4):283–303.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213.

Scott A Crossley. 2020. Linguistic features in writing quality and development: An overview. *Journal of Writing Research*, 11(3):415–443.

Scott A. Crossley, Kristopher Kyle, and Danielle S. McNamara. 2023. Suite of automatic linguistic analysis tools (salat). `https://www.linguisticanalysistools.org/`. Accessed: 2025-04-20.

Mihai Dascalu, Wim Westera, Stefan Ruseti, Stefan Trausan-Matu, and Hub Kurvers. 2017. Readerbench learns dutch: Building a comprehensive automated essay scoring system for dutch language. In *Artificial Intelligence in Education*, pages 52–63, Cham. Springer International Publishing.

Paul Deane. 2013a. Covering the construct: An approach to automated essay scoring motivated by a socio-cognitive framework for defining literacy skills. In M. D. Shermis and J. Burstein, editors, *Handbook of Automated Essay Evaluation*, pages 298–312. Routledge.

Paul Deane. 2013b. On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, 18(1):7–24. Automated Assessment of Writing.

Semire Dikli. 2006. An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, 5(1).

Heejin Do, Yunsu Kim, and Gary Geunbae Lee. 2023. Prompt- and trait relation-aware cross-prompt essay trait scoring. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1538–1551, Toronto, Canada. Association for Computational Linguistics.

Heejin Do, Yunsu Kim, and Gary Geunbae Lee. 2024. Autoregressive score generation for multi-trait essay scoring. *arXiv preprint arXiv:2403.08332*.

Fei Dong, Yue Zhang, and Jie Yang. 2017. Attention-based recurrent convolutional neural network for automatic essay scoring. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 153–162, Vancouver, Canada. Association for Computational Linguistics.

J D'Arcy. 1997. Comparability studies between modular and non-modular syllabuses in gce advanced level biology, english literature and mathematics in the 1996 summer examinations. In *Standing Committee on Research on behalf of the Joint Forum for the GCSE and GCE*.

Mary K Enright and Thomas Quinlan. 2010. Complementing human judgment of essays written by english language learners with e-rater® scoring. *Language Testing*, 27(3):317–334.

Explosion. 2023. Available trained pipelines for dutch: nl_core_news_sm. `https://spacy.io/models/nl#nl_core_news_sm` [Accessed on March 11, 2025)].

Arthur C Graesser, Danielle S McNamara, Max M Louwerse, and Zhiqiang Cai. 2004. Coh-metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, 36(2):193–202.

Ben Hamner, Jaison Morgan, lynnvandev, Mark Shermis, and Tom Vander Ark. 2012. The hewlett foundation: Automated essay scoring. `https://kaggle.com/competitions/asap-aes`. Kaggle.

Yaqiong He, Feng Jiang, Xiaomin Chu, and Peifeng Li. 2022. Automated Chinese essay scoring from multiple traits. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3007–3016, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Sandra Heldsinger and Stephen Humphry. 2010. Using the method of pairwise comparison to obtain reliable teacher assessments. *The Australian Educational Researcher*, 37(2):1–19.

Ian Jones, Marie Bisson, Camilla Gilmore, and Matthew Inglis. 2019. Measuring conceptual understanding in randomised controlled trials: Can comparative judgement help? *British Educational Research Journal*, 45(3):662–680.

Ian Jones and Matthew Inglis. 2015. The problem of assessing problem solving: Can comparative judgement help? *Educational Studies in Mathematics*, 89(3):337–355.

Anders Jonsson. 2014. Rubrics as a way of providing transparency in assessment. *Assessment & Evaluation in Higher Education*, 39(7):840–852.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. `https://github.com/microsoft/LightGBM`. Accessed: 2025-04-22.

Zixuan Ke and Vincent Ng. 2019. Automated essay scoring: A survey of the state of the art. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 6300–6308. International Joint Conferences on Artificial Intelligence Organization.

Kate Kelly, Mary Richardson, and Talia Isaacs. 2022. Critiquing the rationales for using comparative judgement: a call for clarity. *Assessment in Education: Principles, Policy & Practice*, 29:1–15.

Vivekanandan Kumar and David Boulanger. 2020. Explainable automated essay scoring: Deep learning really has pedagogical value. *Frontiers in Education*, 5:572367.

Donald Laming. 2003. *Human judgment: The eye of the beholder*. Cengage Learning, London, United Kingdom.

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.

Sanwoo Lee, Yida Cai, Desong Meng, Ziyang Wang, and Yunfang Wu. 2024. Unleashing large language models' proficiency in zero-shot essay scoring. *arXiv preprint arXiv:2404.04941*.

Marije Lesterhuis, Renske Bouwer, Tine van Daal, Vincent Donche, and Sven De Maeyer. 2022. Validity of comparative judgment scores: How assessors evaluate aspects of text quality when comparing argumentative texts. *Frontiers in Education*, 7:122–131.

Xia Li and Jianda Liu. 2017. Automatic essay scoring based on coh-metrix feature selection for chinese english learners. In *Emerging Technologies for Education*, pages 382–393, Cham. Springer International Publishing.

Andrey Lovakov and Elena R. Agadullina. 2021. Empirically derived guidelines for effect size interpretation in social psychology. *European Journal of Social Psychology*, 51(3):485–504.

R. Duncan Luce. 1959. On the possible psychophysical laws. *Psychological Review*, 66(2):81–95.

H.L.W. Maat, Rogier Kraf, Antal Van den Bosch, Nick Dekker, Maarten Van Gompel, Suzanne Kleijn, and Ko Sloot. 2014. T-scan: A new tool for analyzing dutch text. *Computational Linguistics in the Netherlands*, 4:53–74.

Sandeep Mathias and Pushpak Bhattacharyya. 2020. Can neural networks automatically score essay traits? In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 85–91, Seattle, WA, USA (Online). Association for Computational Linguistics.

Danielle S McNamara, Scott A Crossley, and Philip M McCarthy. 2010. Linguistic features of writing quality. *Written Communication*, 27(1):57–86.

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.

Anneleen V. Mortier, Marije Lesterhuis, Peter Vlerick, and Sven De Maeyer. 2015. Comparative judgment within online assessment: Exploring students feedback reactions. In *Computer Assisted Assessment. Research into E-Assessment*, pages 69–79, Cham. Springer International Publishing.

C Paul Newhouse. 2014. Using digital representations of practical production work for summative assessment. *Assessment in Education: Principles, Policy & Practice*, 21(2):205–220.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Tiffany Potter, Letitia Englund, James Charbonneau, Mark MacLean, Jonathan Newell, and Ido Roll. 2017. Compair: A new online tool using adaptive comparative judgement to support learning with peer feedback. *Teaching & Learning Inquiry*, 5:89.

Dadi Ramesh and Suresh Kumar Sanampudi. 2022. An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, 55(3):2495–2527.

Chaitanya Ramineni and David Williamson. 2018. Understanding mean score differences between the e-rater® automated scoring engine and humans for demographically based groups in the gre® general test. *ETS Research Report Series*, 2018(1):1–31.

Gert Rijlaarsdam, D. Weijen, and Huub Bergh. 1994. Relations between writing processes and text quality: When and how? *Cognition and Instruction*, 12:103–123.

Skipper Seabold and Josef Perktold. 2010. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.

Jeffrey T. Steedle and Steve Ferrara. 2016. Evaluating comparative judgment as an approach to essay scoring. *Applied Measurement in Education*, 29(3):211–223.

Kaveh Taghipour 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*, pages 1882–1891, Austin, Texas. Association for Computational Linguistics.

Tine van Daal, Marije Lesterhuis, Liesje Coertjens, Vincent Donche, and Sven De Maeyer. 2016. Validity of comparative judgement to assess academic writing: examining implications of its holistic character and building on a shared consensus. *Assessment in Education Principles Policy and Practice*, 26:59–74.

San Verhavert, Renske Bouwer, Vincent Donche, and Sven De Maeyer. 2019. A meta-analysis on the reliability of comparative judgement. *Assessment in Education: Principles, Policy & Practice*, 26(5):541–562.

San Verhavert, Sven De Maeyer, Vincent Donche, and Liesje Coertjens. 2018. Scale separation reliability: what does it mean in the context of comparative judgment? *Applied Psychological Measurement*, 42(6):428–445.

Yongjie Wang, Chuang Wang, Ruobing Li, and Hui Lin. 2022. On the use of bert for automated essay scoring: Joint learning of multi-scale essay representation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3416–3425, Seattle, United States. Association for Computational Linguistics.

A Rubric Description

Main Component	**Rubric**	**Description**
Structure	Title	The text has a title that clearly matches the content of the text.
Structure	Construction	The text contains a clear division into: introduction, argumentation, and conclusion.
Structure	Layout	The text is well-organized. There is a clear division into paragraphs. Paragraphs are separated by: blank lines, indentation, or starting on a new line.
Structure	Subtopic	Each paragraph has its own single (sub)topic.
Structure	Relationships between Paragraphs	There is a clear 'train of thought' between paragraphs: based on the text, coherence relationships between paragraphs can be clearly (easily) identified.
Structure	Continuity	Information that belongs together is also grouped together in the text.
Content	Introduction	In the introduction, the proposition/statement is presented, and optionally, the writer's opinion on the proposition is also made clear.
Content	Persuasion	It is clear what the writer wants to convince the reader of: a choice for or against the presented proposition.
Content	References	The text contains at least two (parts of) references, which are meaningfully incorporated into the text. For example, they support the argumentation or are used as an example in the introduction.
Content	Citations (quoting from references)	The quotes from the references are correctly marked in the text. Direct quotes (between quotation marks) and paraphrases both have a source citation.
Content	Reader Focus	The text is easily understandable for a reader unfamiliar with the assignment. For example, there is no reference to the writing task assignment or the writer's environment.
Content	Reader Engagement	The reader is clearly engaged with the text through examples referring to daily life or common experiences.
Content	Conclusion	The text contains a clear conclusion that aligns with the rest of the text and from which the writer's opinion is evident. It is clear that this concludes the text.
Argumentation	Support	The argumentation consists of multiple arguments that support the writer's opinion.
Argumentation	Relevance	The argumentation does not contain too much superfluous information, i.e., information that does not contribute to supporting the writer's opinion.
Argumentation	Indication of Argumentation	The arguments are clearly recognizable as arguments; e.g., through the use of constructions like "therefore I believe (do not believe) that...", "I find/think...", "I (do not) agree with this", etc.
Argumentation	Referential and Coherence Relations	The referential and coherence relations are clear when implicit, or explicitly marked. Examples of markers are: therefore, thereby, thus, because, since, first, second, third, then, etc.
Language	Grammar and Spelling	The text contains no grammatical and/or spelling errors.
Language	Punctuation	Punctuation marks are applied correctly.

Language	Style	The tone and word choice are appropriate for the purpose and audience of the text.

Table 4: List of rubrics used to assess Dutch essays on argumentative writing. Each rubric was assigned a score between 1 and 5. For the original Dutch version, see Coertjens et al. (2017).

B T-Scan Configuration

Parameter	Value
Overlap Size	50
Frequency Clipping	99.0
MTLD factor size	0.72
Use Alpino parser?	yes
Store Alpino output?	yes
Use Wopr?	yes
One sentence per line?	no
Prevalence data	Belgium
Word Frequency List	subtlex_words.freq
Lemma Frequency List	subtlex_lemma.freq
Top Frequency List	subtlex_words20000.freq
Compound split method	compound-splitter-nl

Table 5: Configuration of T-Scan (Maat et al., 2014) used to extract linguistic features from Dutch essays.

C Hyperparameters

Rubric	Model	Optimal Hyperparameters
Title	LightGBM	boosting_type = gbdt, num_leaves=31, max_depth=-1, learning_rate=0.1, n_estimators=100, min_child_weight=0.001, min_child_samples=20
Construction	XGBoost	colsample_bytree = 0.6, gamma = 0.1, learning_rate = 0.05, max_depth = 10, min_child_weight = 7, n_estimators = 800, reg_alpha = 0.5, reg_lambda = 0.5, subsample = 0.4
Layout	RandomForest	max_depth = None, min_samples_split = 2, n_estimators = 100
Subtopic	XGBoost	colsample_bytree = 1.0, gamma = 0, learning_rate = 0.01, max_depth = 20, min_child_weight = 7, n_estimators = 300, reg_alpha = 0.0, reg_lambda = 0.0, subsample = 0.6
Relationships	ElasticNet	alpha = 0.01, l1_ratio = 0.3
Continuity	RandomForest	max_depth = None, min_samples_split = 2, n_estimators = 100
Introduction	ElasticNet	alpha = 0.01, l1_ratio = 0.3
Persuasion	ElasticNet	alpha = 0.01, l1_ratio = 0.1
References	Lasso	alpha = 0.001
Citations	XGBoost	colsample_bytree = 0.6, gamma = 0.1, learning_rate = 0.05, max_depth = 10, min_child_weight = 7, n_estimators = 800, reg_alpha = 0.5, reg_lambda = 0.5, subsample = 0.4
Reader Focus	ElasticNet	alpha = 0.01, l1_ratio = 0.1
Reader Engagement	ElasticNet	alpha = 0.01, l1_ratio = 0.1
Conclusie	XGBoost	colsample_bytree = 1.0, gamma = 0, learning_rate = 0.01, max_depth = 20, min_child_weight = 7, n_estimators = 300, reg_alpha = 0.0, reg_lambda = 0.0, subsample = 0.6
Support	Lasso	alpha = 0.001
Relevance	ElasticNet	alpha = 0.01, l1_ratio = 0.1
Indication	ElasticNet	alpha = 0.01, l1_ratio = 0.1
Reference Cohesion Relationships	XGBoost	colsample_bytree = 1.0, gamma = 0, learning_rate = 0.01, max_depth = 20, min_child_weight = 7, n_estimators = 300, reg_alpha = 0.0, reg_lambda = 0.0, subsample = 0.6
Grammar and Spelling	Lasso	alpha = 0.001
Punctuation	ElasticNet	alpha = 0.01, l1_ratio = 0.1
Style	ElasticNet	alpha = 0.01, l1_ratio = 0.1

Table 6: The optimal hyperparameters that were selected for the best-performing model during LOOCV. For each run of LOOCV, the optimal hyperparameters were selected based on the lowest average MAE, using 20-fold cross-validation with 100 randomized iterations. For brevity, we only report the most frequently selected optimal hyperparameters for the best models.