Through the Sentence Lens: Explainable Essay Scoring through Fine-Grained Predictions

Daniel Mora[1,2], Stefan Keller[3], Andrea Horbach[1,2]

[1]Leibniz Institute for Science and Mathematics Education, Kiel, Germany
[2]Kiel University, Kiel, Germany,
[3]Zurich University of Teacher Education, Zurich, Switzerland
Correspondence: mora@leibniz-ipn.de

Abstract

Beyond performance, model transparency is a crucial factor in Automated Essay Scoring, yet current systems often lack explainability, limiting their pedagogical value and users' trust. Existing explainability methods, such as gradient-based attribution or feature-importance approaches, either produce counter-intuitive explanations or are too complex for classroom use. To address this limitation, we make use of fine-grained prediction at the sentence level as a way to enhance explainability. We propose ablation strategies to derive sentence-level pseudo scores from essay-level gold scores and use them to train sentence-level models. We evaluate their performance against essay-level baselines on two datasets (ASAP and MEWS), and compare their sentence-level output to a human baseline. Results indicate a trade-off between essay-level performance and sentence-level granularity. For the *language quality* trait, most sentence-level models achieve performance comparable to the essay-level baseline, whereas for *content*, the approach yields more positive results on prompts with shorter student texts.

1 Introduction

Automated Essay Scoring (AES) systems are tools designed to evaluate free-text written responses by assigning a score typically using machine learning algorithms. Despite their significant benefits for teachers, such as reducing the workload and making it possible to provide timely feedback to students, their implementation in real-life educational scenarios remains limited due to the lack of explainability, that is, the model's reasoning cannot be made understandable to humans (Li et al., 2022). Whether used for formative feedback or in high-stakes exams, improving such transparency for end users is essential to enhance the overall usability of AES systems (Hall et al., 2024).

Most approaches to improving model explainability have focussed on developing inherently interpretable models or applying post-hoc local explanation methods. However, these techniques often lack pedagogical value in educational contexts: gradient-based attribution methods can be counterintuitive (Parekh et al., 2020), and feature-importance approaches (Kumar and Boulanger, 2020; Urrutia et al., 2025) frequently suffer from construct validity issues (Beigman Klebanov and Madnani, 2020). Rubric-based or fine-grained predictions could offer more intuitive explanations for teachers and students, but their development is constrained by the scarcity of annotated datasets (Ramesh and Sanampudi, 2022), which limits the training of machine learning models.

To address this gap in explainability, we build on the idea of fine-grained prediction by focusing on scoring each trait at the sentence level, rather than assigning only an overall holistic score. Because there is a lack of sentence-level annotations, we propose a method for deriving sentence-level pseudo-scores from essay-level annotations using an ablation strategy. Specifically, we use a model trained at the essay-level to predict the relevance of a sentence by comparing the difference between the prediction on the full essay and the prediction when that particular sentence is missing. This difference in prediction indicates whether the sentence contributes positively or negatively to the overall score.

As a next step, we train sentence-level models using the pseudo scores we had derived. Table 1 illustrates an example output showing sentence-level predictions that can be useful for learners' writing, as receiving feedback on individual sentences helps them focus their efforts when revising their texts. Finally, we evaluate the proposed approach by comparing the performance of sentence-level models against essay-level baselines and by evaluating the automated sentence-level scores against

Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications, pages 221–233

human annotations[1].

This research thus makes two contributions:

- **Model Performance and Ablation Impact:** We compare the performance of different AES models architectures trained at the essay level using annotated gold scores with models trained at the sentence level using pseudo scores generated through ablation strategies on ASAP and MEWS datasets.
- **Alignment with Human Judgements:** We further examine the extent to which the sentence rankings predicted by sentence-level models align with human evaluations of sentence quality on a subset of the MEWS dataset.

Text	Score
Older children, like with the age of five or six or elder, it is of course right and possible to watch television with a time, which is given by the parents, so that the children can watch a serie for one hour and then it's finished for this day.	4.1
So all in all, I agree at some parts with the statement, but not at all.	1.4
The main thing, that is important is the bad influence by television for too young children, so it could exists problems in future, which are developed by the television.	0.9

Table 1: Sentence-level predictions for the last three sentences of a random essay in the MEWS dataset (Keller et al., 2020). Score range is 0–5.

2 Related Work

Research on transparency in AES has taken several directions, focusing on model interpretability, model explanations, trait scoring, or fine-grained predictions.

2.1 Model Interpretability

Model interpretability refers to explaining the intrinsic properties of the model that make it understandable to humans (Doshi-Velez and Kim, 2017). As the computational complexity of a model increases, however, its inherent interpretability decreases (Barceló et al., 2020; Bexte et al., 2024). For example, in a linear regression model, the prediction is expressed as a weighted sum of features, where each coefficient reflects the relevance of a feature (Molnar, 2025).

[1]Source code available at https://github.com/melanchthon19/sentence-level-aes

In this context, Urrutia et al. (2025), Kumar et al. (2019), and Eltanbouly et al. (2025) use linear regressor and Random Forest models due to their higher degree of intrinsic interpretability and provide the most relevant features as feedback to the learner. Although complex models tend to outperform simpler ones, this is not always the case (Li and Ng, 2024), and simpler models should be preferred if interpretability is a main concern (Rudin, 2019).

The use of generative LLMs in the AES context has been widely explored, yielding mixed results (Mizumoto and Eguchi, 2023; Lee et al., 2024; Mansour et al., 2024; Yoshida, 2025). Even though generative models are able to produce explanations using natural language when instructed to provide insights into their decision-making process, these explanations could not reflect the model's underlying reasoning, and can be yet misleading (Turpin et al., 2023). Methods such as chain-of-thought prompting and self-explanations have been proposed to improve explainability, but they remain limited, as they do not provide direct access to the internal decision process of the model (Lanham et al., 2023; Dehghanighobadi et al., 2025).

2.2 Model Explanations

If a model is too complex, an interpretation algorithm (Li et al., 2022) can be applied to reveal how the model makes decisions, usually in the form of a post-hoc local explanation (Madsen et al., 2022). Common approaches for explaining neural networks are attention-based (Dong et al., 2017; Yang et al., 2020), gradient-based (Alikaniotis et al., 2016), and feature-based attribution methods (Kumar and Boulanger, 2020).

Although post-hoc local explanations can be useful, their pedagogical value remains limited. For example, gradient-based attributions may highlight factors that do not align with human intuitions, such as different sub-tokens of the same word contributing to opposite predictions (Chefer et al., 2021), or the order and even the presence of certain words being irrelevant to the model (Parekh et al., 2020). Similarly, in feature-based attribution methods, the identified feature may be only an indirect proxy that is not explicitly assessed, such as essay length (Jeon and Strube, 2021), or it may be so heavily engineered that the model loses decomposability (Lipton, 2018).

2.3 Trait Scoring

Trait or rubric-based scoring disaggregates a holistic score into specific traits defined in a rubric. Mathias and Bhattacharyya (2018) enrich the ASAP dataset with trait-level annotations, and Yoo et al. (2025) extend this line of work by exploring augmentation strategies. Other studies focus on modelling approaches: Mizumoto et al. (2019) propose a BiLSTM model combined with supervised attention for justification identification; Kumar and Boulanger (2021) examine the relationship between groups of linguistic features and specific traits, as well as their contribution to holistic scoring; Lohmann et al. (2024) compare models and input features on trait scoring; and Eltanbouly et al. (2025) leverage generative LLMs to extract rubric-based features. While trait scoring provides more explanations than a single holistic score, the assigned scores remain at the essay level, with no finer-grained evaluation.

2.4 Sentence-Level Predictions

In addition to trait-level scoring, explainability can also be enhanced through fine-grained prediction, where models score sentences instead of providing just one overall score. Because manual annotations at the sentence level are scarce and expensive, most supervised approaches must circumvent the lack of training data.

Andersen et al. (2013) frame the task as discriminative ranking and compare two approaches. The first uses an essay-level model to rank sentences, which produces weak results. The second derives pseudo sentence scores by dividing holistic essay scores by the number of manually annotated errors per sentence, which yields stronger sentence-ranking performance. When aggregating back to the essay level, however, the sentence-level model underperforms compared to the model trained directly on holistic essay scores. No direct evaluation of the sentence rankings is reported.

Woods et al. (2017) propose an ablation strategy that avoids fine-grained sentence-level annotations. Using an essay-level model, they estimate sentence contributions by omitting sentences and measuring changes in predicted scores. To account for length bias, essay length is included as a fixed feature, and score changes are evaluated against a background distribution of unannotated essays. Results show a slight advantage over a baseline in identifying trait-relevant sentences, though no direct comparison is made with the original essay-level model.

Hossain and Mustafa (2023) explore the use of an adjacent sentence attention mechanism designed to capture contextual dependencies between neighbouring (preceding and following) sentences. Their model underperforms compared to pretrained baselines, mainly due to the scarcity of training data.

These approaches have some limitations: they either do not directly evaluate the performance of the sentence-level model, lack comparison with essay-level models trained on manual annotations, or are constrained by small datasets. This study addresses these gaps by training sentence-level models using pseudo scores, evaluating their performance against essay-level baselines, and validating sentence-level outputs through a human annotation study.

3 Datasets

We use the ASAP[2] dataset and its trait-enriched extension ASAP++ (Mathias and Bhattacharyya, 2018), as well as the MEWS (Keller et al., 2020, 2024) dataset. Both corpora cover a wide range of argumentative and narrative writing tasks, and represent diverse student populations. ASAP consists of essays written by native English students from grades 7 to 10 for 8 different prompts, whereas MEWS represents English L2 learners from German and Swiss upper secondary schools with German as L1 for two prompts (TE and AD). In our experiments we also considered their combination (TEAD). Further details of the datasets are provided in the original articles.

For consistency across datasets, we select prompts that include comparable trait-level annotations. We also exclude document-level traits such as *holistic* and *organisation* given that evaluating their quality goes beyond the sentence unit.

The final set of traits used in our experiments include *content* and *language quality*. According to the evaluation rubrics, these traits include effective language command, accurate syntactic and idiomatic usage, sufficient and engaging details, relevant explanations, and well-supported ideas. Table 2 summarises the score distribution for the two traits across prompts, which follow an approximately bell-shaped distribution.

[2]https://www.kaggle.com/datasets/lburleigh/asap-2-0

Prompt	Count	Lang. Qual.	Content
MEWS			
AD	1169	2.79 ± 1.10	3.35 ± 0.96
TE	1099	3.12 ± 1.18	3.68 ± 0.84
ASAP			
1	1783	–	3.85 ± 0.99
2	1801	–	3.22 ± 1.17
3	1726	1.47 ± 0.85	1.43 ± 0.84
4	1770	1.06 ± 0.88	1.11 ± 0.97
5	1805	2.24 ± 0.94	1.88 ± 1.00
6	1800	2.05 ± 0.92	1.85 ± 1.11
7	1569	–	1.84 ± 0.80
8	723	–	3.80 ± 0.63

Table 2: Label distribution of *language quality* and *content* (mean and SD) per prompt for MEWS and ASAP.

4 Experimental Study: Sentence-Level Predictions

In order to evaluate if fine-grained essay scoring improves explainability while preserving the scoring performance, we (i) train essay-level models on the manually annotated scores, (ii) generate sentence-level pseudo scores through ablation strategies, (iii) train sentence-level models using these pseudo scores, and (iv) evaluate the aggregated essay-level predictions against the human-annotated gold standard using Quadratic Weighted Kappa (QWK) (Cohen, 1968). Figure 1 depicts the overall process.

4.1 Models

We train essay-level models representing the main AES paradigms (Xu et al., 2024): linguistic feature-based, transformer-based, and hybrid models. The corresponding architectures are shown in Figure 2 and training hyperparameters in Appendix A.1. Sentence-level models use the same architectures and hyperparameters as their corresponding essay-level baselines.

Linear Regressor (LR): This model is trained on 220 handcrafted linguistic features extracted using the tool described in Lohmann et al. (2024). The features include lexical, syntactic, cohesion-related, error-based, frequency-based, and length-based indicators. The features are scaled and multicollinear features are removed before being fed into the model.

DistilBERT (LLM): A DistilBERT[3] model (Sanh et al., 2019) with pretrained weights is used to encode the input essay. The essay representation is obtained by applying max pooling to the final hidden layer, and it is then passed to a custom regression head composed of two linear layers, which is fine-tuned on top of the encoder.

Hybrid Model (Hybrid): This model integrates the first two approaches by combining the max pooled representation of the final hidden layer from the DistilBERT encoder with a feedforward network that processes the 220 linguistic features used in the linear regression model. The two representations are concatenated and passed through a shared regression head composed of two linear layers.

4.2 Pseudo Scores via Ablation

The main intuition is that the quality of an essay is not equal across all sentences, and that some sentences have a stronger influence on a positive or negative score. Removing a sentence should cause the model to change its prediction according to the relevance of the sentence.

Using the model trained at the essay level, we estimate the magnitude and direction of this relevance by computing the difference in predictions on ablated (missing one or more sentences) and full essays. If the prediction of the ablated essay is higher compared to that of the full essay, it means that the missing fragment (one or more sentences) was lowering the score. Therefore, this fragment should be assigned a lower pseudo score compared to the score assigned at the essay level.

Once pseudo scores are generated, they are normalised to match the target scoring range. Normalisation proceeds in three steps. Z-score normalisation centres the values at zero and scales them to have a unit standard deviation. Tanh non-linear transformation then compresses outliers beyond one standard deviation, and, finally, min–max scaling maps the transformed values to the target scoring range. Figure 6 in Appendix A.2 illustrates the resulting distribution of pseudo scores after normalisation.

4.2.1 Sentence Ablation

To generate sentence-level pseudo scores, each essay is first segmented into individual sentences using spaCy (Honnibal et al., 2020). Then, the ablation strategy follows the leave-one-out method by removing one or more sentences at a time, predicting a score on the ablated essay, and computing a pseudo score for that particular fragment using the following formula:

$$\hat{y} = y - \alpha \cdot \left(\hat{y}_{\text{loo}} - \hat{y}_{\text{essay}} - b\right), \qquad (1)$$

where $\hat{y}$ is the pseudo score, y is the true essay-level score, $\hat{y}_{\text{loo}}$ is the prediction without the target

[3] distilbert-base-uncased

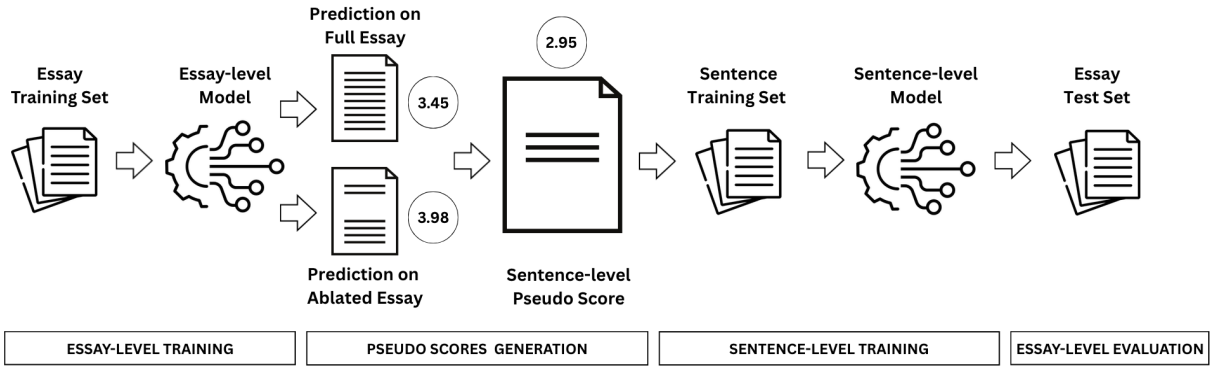

Figure 1: Pipeline for training sentence-level models using pseudo sentence scores derived from essay-level annotations via ablation. The pseudo scores generation is described in Section 4.2.

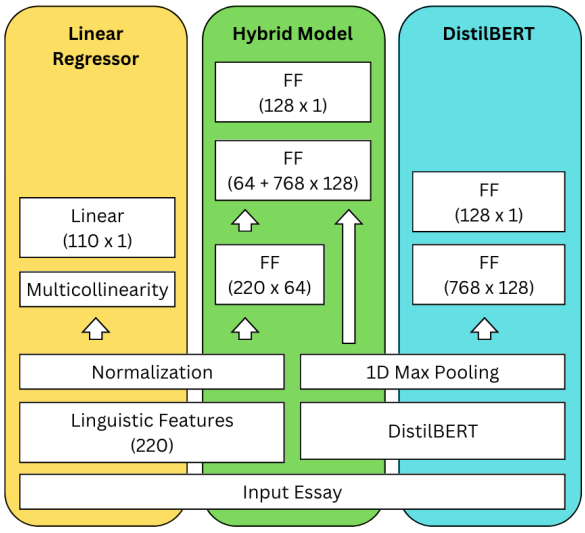

Figure 2: Model architectures used for training both essay- and sentence-level models.

fragment (leave-one-out), $\hat{y}_{\text{essay}}$ is the prediction on the full essay, and b is a bias term estimated as the mean difference between $\hat{y}_{\text{loo}}$ and $\hat{y}_{\text{essay}}$ across the entire dataset, which accounts for the model's tendency to prefer longer inputs with more context. A scaling factor α amplifies the contribution difference, given that the model is generally stable on ablated essays, and its predictions have a low variance within the same essay. For the experiments, we varied α across the values 1, 10, 20, and 30.

For multi-sentence ablation, a sliding window n is used for leaving out several contiguous sentences when making a prediction. The contribution of a single sentence is then computed as the average contribution across all windows in which the sentence appears. For the experiments, the window size parameters n was varied between 3 and 5.

4.3 Models' Performance Results

Figure 3 depicts the performance of the best sentence-level models compared to their essay-level baseline across traits and prompts. The primary benefit of this comparison is to highlight the setups in which the sentence-level model trained on pseudo-scores matches the performance of the essay-level model trained on annotated gold scores while offering greater granularity.

4.3.1 Essay-Level Results

Across the three models, the ASAP dataset yields higher and more stable results (mean QWK 0.61 ± 0.08) than MEWS (0.50 ± 0.18). Across both datasets, LR performs best (average QWK 0.63), achieving the best score in 67% of cases, followed by Hybrid (0.59; best in 28% of cases) and LLM (0.50). The embeddings alone used by the LLM are not sufficient to model both *content* and *language quality* across the two datasets, compared to the other models that included linguistic features.

Comparing LR and Hybrid models, concatenating embeddings with linguistic features did not consistently improve performance, with gains observed only for three prompts. Essay baseline results suggest that using simpler models preserves performance while retaining intrinsic interpretability, as also demonstrated in Li and Ng (2024).

4.3.2 Sentence-Level Results

Language Quality: Overall, sentence-level models show similar performance compared to their essay-level baselines. The LLM model has higher QWK variance compared to LR and Hybrid models. The LLM obtains similar results on two prompts, decreases by -0.21 on average on four prompts, and increases by +0.17 on one prompt. The LR and Hybrid model achieve comparable performance on

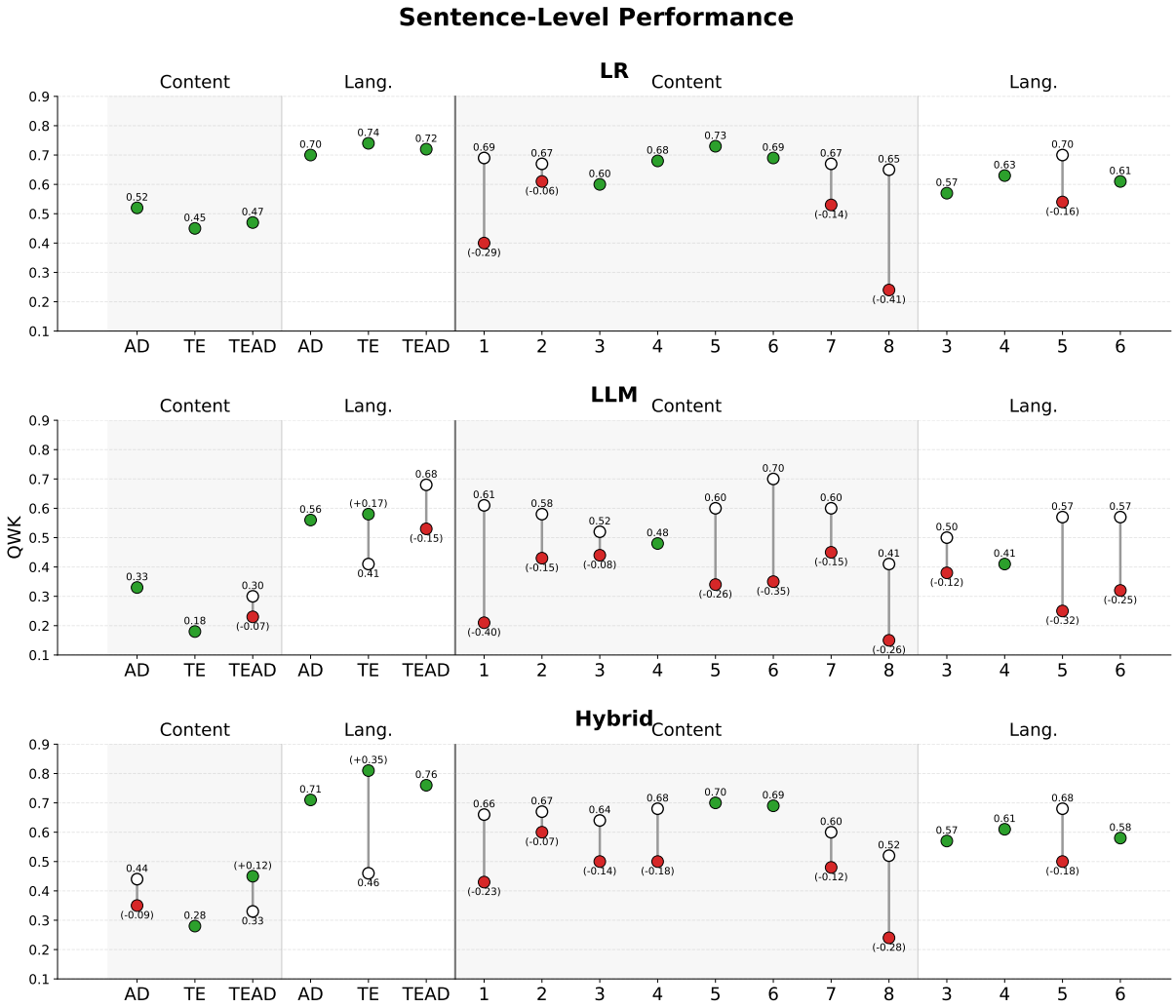

Figure 3: Performance (QWK) of the essay-level baseline (white dot) and the best sentence-level model on the ASAP and MEWS datasets per prompt and trait. All differences are statistically significant based on 95% bootstrap confidence intervals. Full results are in Appendix B.

all prompts, except on prompt 5 (ASAP), in which there is a decline of -0.17 points on average. The highest increase of +0.35 is obtained by the Hybrid model for the TE prompt (MEWS).

Figure 4 presents the confusion matrices for the essay- and sentence-level hybrid models trained on the TE prompt. The essay-level model predicts scores clustered around the middle of the scale, with a narrow range (1.5–4.0), often assigning a score of 3.0 to essays whose true scores span 1.0–5.5. In contrast, the sentence-level model, benefitting from more distributed pseudo scores, covers a wider range (1.0–5.5), and its predictions of 3.0 align more closely with true scores between 2.0 and 4.0.

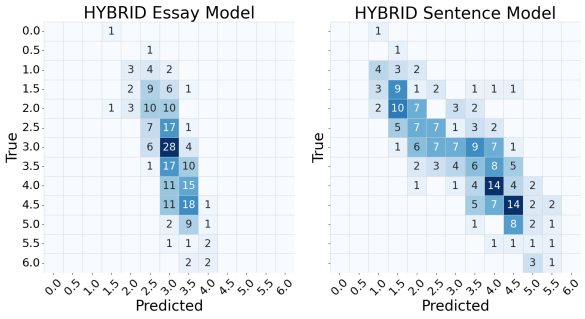

Figure 4: Confusion matrices of the Hybrid essay- and best sentence-level models trained on the TE prompt for the *language quality* trait.

Content: Results vary depending on the dataset and prompt. For the MEWS dataset, models show similar performance compared to their essay-level baselines. Exceptions include the LLM model on the TEAD prompt and the Hybrid model on the AD prompt, both showing a decrease of less than

one point. The Hybrid model, on the contrary, shows improvement of +0.12 when both prompts are combined (TEAD).

For the ASAP dataset, models obtained a lower performance on prompts 1, 2, 7, and 8 for all three models. The LR model, however, maintains comparable performance compared to its essay-level baseline on prompts 3 to 6.

This difference in performance for the *content* trait seems to be primarily related to variation in essay length across prompts. Prompts 3 to 6 that have similar results between essay- and sentence-level models include both source-dependent and independent tasks, but the student essays for these prompts are shorter on average than those for prompts 1, 2, 7, and 8.

Ablation Strategy: A comparison of single versus multi-ablation strategies reveals that the best performances across traits and prompt setups are achieved in 69% of cases with single ablation and in 82% with an alpha value of one. The LLM model is the only one to improve further with higher alpha values, likely because its ablated predictions remain nearly constant across ablated essays, with a variance closer to 0.1. Therefore, amplifying these differences enhances the model's ability to discriminate between sentences. For the LR and Hybrid models that use linguistic features, removing one or more sentences causes a substantial shift in the input features. This shift is sufficient to determine both the relevance and polarity of a given sentence during the ablation process, which explains why an alpha value of one was the most effective setting for these two models. The window size of 5 was beneficial to some extent (19% of cases) for all three models, with higher improvements for the *content* trait.

Sentence-level results indicate that the proposed approach can provide more detailed and informative feedback without sacrificing performance for some of the prompts and traits. This granularity allows for clearer differentiation between essays of varying quality and offers more actionable insights for teaching and learning.

5 Annotation Study: Sentence-Level Language Quality Scores

In addition to model performance comparisons, we also conduct a human annotation study. The goal is to assess whether the proposed methods agree with human judgement on identifying sentences that contribute most positively or most negatively to the overall essay quality, focusing on *language quality* as the main criterion as it is more easily identifiable at the sentence level.

5.1 Annotation Task

To this end, five expert evaluators independently annotated a stratified random sample of 100 essays drawn from the MEWS test set. The sample was filtered according to essay length in terms of number of sentences (10–15) and character count (1,000–2,000). We manually screened the essays and excluded one partially written in a language other than English. The final sample includes 14 low-, 70 mid- and 15 high-scoring essays, evenly distributed across prompts (54 from TE and 45 from AD), resulting in 99 essays in total.

The annotators selected three sentences with the most positive impact and three with the most negative impact on the overall essay based on language command, syntactic variety, and idiomatic usage. These core aspects define the *language quality* trait as specified in the evaluation rubric.

5.2 Annotation Results

Each annotation is mapped to a numerical value (positive = 1, negative = –1, neutral = 0) and summed up across annotators. A given sentence can receive an overall score between -5 and +5, where extreme values indicate complete agreement and zero represents neutral annotations or multiple annotations balanced out. Table 3 shows an example of annotated sentences.

Text	+	-	=
That is the reason why I think, it's impossible being nasty, if you are a friendly person normally.	0	4	-4
Another reason, why I think it's better to have a good relation between students and teachers, is that students often learn about the behaviour of their teachers.	2	1	+1
So if a teacher wants them to learn about a subject as best without any kind of happiness or confidence, the students couldn't learn about patience and any kind of stuff which they have to know in their future, like being comfortable and friendly.	1	4	-3

Table 3: Human annotations on three consecutive sentences from a randomly sampled essay regarding their positive (+) or negative (-) contribution to the overall essay considering *language quality* as the criterion.

The distribution of aggregated scores is shown in Figure 5: almost 20% of sentences are consid-

ered neutral (10% without any annotations and another 10% balanced out after aggregating conflicting contributions), there is higher agreement among positive sentences (score of 4 or 5), and a greater proportion of negative sentences with lower agreement (score of -1 or -2). Roughly 15% of sentences show high agreement, reaching an aggregated score greater than 4 in both the positive and negative directions. This pattern is also reflected in inter-annotator agreement measures: Krippendorff's alpha score is 0.17 and pairwise inter-annotator Cohen's kappa ranges from 0.14 to 0.22.

5.3 Validation of the Sentence Level: Human–Model Correlation

We evaluate the alignment between human annotations and model outputs by computing the Spearman correlation between the aggregated expert scores and the model outputs, including both the generated pseudo scores per model and sentence-level predictions. Table 4 reports the average correlation across essays and the corresponding standard deviation for both model outputs.

There is a modest Spearman correlation between the LR, LLM, and Hybrid models trained on both prompts (TEAD), based on their pseudo scores and predictions, compared with the aggregated annotation scores. The LLM model achieves the highest averaged correlation (0.29), followed by the Hybrid model (0.17), whereas the LR model shows a negative correlation (–0.15). Although the correlation is weak overall, all models nevertheless capture a meaningful underlying pattern compared to a random baseline.

The positive correlations obtained by the LLM model suggest that its sentence-level outputs are more aligned with human annotations, even though averaging sentence scores leads to lower QWKs as seen in Figure 3. This indicates that this model is better able to capture the lexical and syntactic patterns that annotators consider when judging the *language quality* of a sentence, which is likely due to the use of word embeddings. The Hybrid model also shows a meaningful correlation, though weaker, suggesting that the inclusion of linguistic features do not substantially enhance the model's ability to identify the most positive and most negative sentences, but instead add noise.

Although the best sentence-level LR model has a similar performance compared to its essay-level baseline on 13 out of the 18 different prompts, as shown in Figure 3, its sentence-level predictions diverge from human annotations with a negative correlation. This indicates that, even though aggregated sentence-level predictions correlate with overall essay scores, the individual predictions do not reflect human judgements of which sentences contribute positively or negatively. This difference can be explained by the fact that linguistic features at the sentence level are less informative, as many rely on frequencies, ratios, and distributions that require broader context.

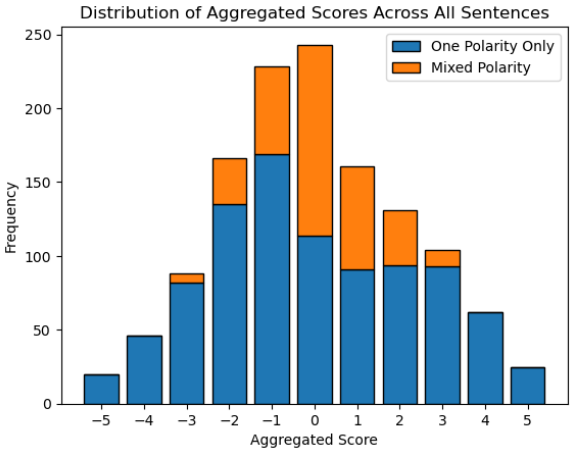

Figure 5: Distribution of aggregated sentence-level human annotations. ±5 indicates complete agreement among all five annotators on sentence polarity.

Model	Predictions	Pseudo Scores
LLM	0.30 (0.26)	0.11 (0.25)
Hybrid	0.19 (0.28)	0.14 (0.29)
LR	-0.15 (0.28)	0.16 (0.30)

Table 4: Spearman correlations (mean and SD) between aggregated human annotations at the sentence level and models' output (predictions and pseudo scores).

6 Conclusion

In this work, we explore fine-grained prediction as a means of enhancing the explainability of AES models. To address the lack of manual sentence-level annotations, we propose ablation strategies to generate pseudo scores for training sentence-level models and we evaluate their effectiveness against essay-level baselines using ASAP and MEWS datasets. We further examine how the outputs of these models correlate with human sentence-level annotations of the *language quality* trait on a subset of the MEWS dataset.

Overall, there is a trade-off between essay-level performance and sentence-level granularity, though it varies depending on the specific trait and ablation setup. The proposed approach is more effective for the *language quality* trait than for *content*, and for *content*, it performs better on shorter texts. Both model outputs and human judgements show relatively weak correlations, highlighting the subjective nature of the task. Despite this, averaging predictions across sentences yields strong essay-level results even when the underlying sentence-level predictions do not closely align with human reasoning. This points to the need for alternative aggregation strategies, such as assigning different weights to sentences rather than treating them all equally when averaging.

Our experiments also show that simpler models, such as a linear regressor, are able to produce competitive results at a lower computational cost and with a higher degree of intrinsic interpretability. This highlights that the trade-off between complexity and performance does not always hold, and that it might also be preferred in educational settings where transparency is crucial. Applying further explainability methods to sentence-level models would offer a clearer understanding of their internal reasoning and enhance model transparency.

Limitations

This study is limited to traits that are conveyed within the span of a sentence, such as *content* and *language quality*. The ASAP and MEWS datasets consist of English texts, and results may differ for other languages. Different encoders for the LLM and Hybrid model might have an impact on performance.

7 Ethical Considerations

All data annotation procedures were conducted by qualified annotators who were compensated in accordance with legal standards. The models were trained using data representatives of the population from which the data were obtained. Consequently, any application of these models to new or distinct populations should be undertaken with appropriate measures to ensure fairness, validity, and the minimisation of potential biases.

8 Bibliographical References

References

Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. 2016. Automatic text scoring using neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 715–725, Berlin, Germany. Association for Computational Linguistics.

Øistein E. Andersen, Helen Yannakoudakis, Fiona Barker, and Tim Parish. 2013. Developing and testing a self-assessment and tutoring system. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 32–41, Atlanta, Georgia. Association for Computational Linguistics.

Pablo Barceló, Mikaël Monet, Jorge Pérez, and Bernardo Subercaseaux. 2020. Model interpretability through the lens of computational complexity. In *Advances in Neural Information Processing Systems*, volume 33, pages 15487–15498. Curran Associates, Inc.

Beata Beigman Klebanov and Nitin Madnani. 2020. Automated evaluation of writing – 50 years and counting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7796–7810, Online. Association for Computational Linguistics.

Marie Bexte, Andrea Horbach, and Torsten Zesch. 2024. Strengths and weaknesses of automated scoring of free-text student answers. *Informatik Spektrum*, 47(3):78–86.

Hila Chefer, Shir Gur, and Lior Wolf. 2021. Transformer interpretability beyond attention visualization. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 782–791.

J Cohen. 1968. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213—220.

Zahra Dehghanighobadi, Asja Fischer, and Muhammad Bilal Zafar. 2025. Can LLMs Explain Themselves Counterfactually? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7787–7815, Suzhou, China. Association for Computational Linguistics.

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.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *Preprint*, arXiv:1702.08608.

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, Vienna, Austria. Association for Computational Linguistics.

Erin Hall, Mohammed Seyam, and Daniel Dunlap. 2024. Exploring explainability and transparency in automated essay scoring systems: A user-centered evaluation. In *Learning and Collaboration Technologies*, pages 266–282, Cham. Springer Nature Switzerland.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python.

Mehadi Hossain and Hossen A Mustafa. 2023. Automated writing evaluation using sentence by sentence scoring model. In *2023 International Conference on Next-Generation Computing, IoT and Machine Learning (NCIM)*, pages 1–6.

Sungho Jeon and Michael Strube. 2021. Countering the influence of essay length in neural essay scoring. In *Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing*, pages 32–38, Virtual. Association for Computational Linguistics.

Stefan D. Keller, Johanna Fleckenstein, Maleika Krüger, Olaf Köller, and André A. Rupp. 2020. English writing skills of students in upper secondary education: Results from an empirical study in switzerland and germany. *Journal of Second Language Writing*, 48:100700.

Stefan D. Keller, Julian Lohmann, Ruth Trüb, Johanna Fleckenstein, Jennifer Meyer, Thorben Jansen, and Jens Möller. 2024. Language quality, content, structure: What analytic ratings tell us about efl writing skills at upper secondary school level in germany and switzerland. *Journal of Second Language Writing*, 65:101129.

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

Vivekanandan S. Kumar and David Boulanger. 2021. Automated essay scoring and the deep learning black box: How are rubric scores determined? *International Journal of Artificial Intelligence in Education*, 31(3):538–584.

Yaman Kumar, Swati Aggarwal, Debanjan Mahata, Rajiv Ratn Shah, Ponnurangam Kumaraguru, and Roger Zimmermann. 2019. Get it scored using autosas — an automated system for scoring short answers. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, and 11 others. 2023. Measuring Faithfulness in Chain-of-Thought Reasoning. *Preprint*, arXiv:2307.13702.

Gyeong-Geon Lee, Ehsan Latif, Xuansheng Wu, Ninghao Liu, and Xiaoming Zhai. 2024. Applying large language models and chain-of-thought for automatic scoring. *Computers and Education: Artificial Intelligence*, 6:100213.

Shengjie Li and Vincent Ng. 2024. 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 (Volume 1: Long Papers)*, pages 7661–7681, Bangkok, Thailand. Association for Computational Linguistics.

Xuhong Li, Haoyi Xiong, Xingjian Li, Xuanyu Wu, Xiao Zhang, Ji Liu, Jiang Bian, and Dejing Dou. 2022. Interpretable deep learning: interpretation, interpretability, trustworthiness, and beyond. *Knowledge and Information Systems*, 64(12):3197–3234.

Zachary C. Lipton. 2018. The mythos of model interpretability. *Commun. ACM*, 61(10):36–43.

Julian F. Lohmann, Fynn Junge, Jens Möller, Johanna Fleckenstein, Ruth Trüb, Stefan Keller, Thorben Jansen, and Andrea Horbach. 2024. Neural networks or linguistic features? - comparing different machine-learning approaches for automated assessment of text quality traits among l1- and l2-learners' argumentative essays. *International Journal of Artificial Intelligence in Education*.

Andreas Madsen, Siva Reddy, and Sarath Chandar. 2022. Post-hoc interpretability for neural nlp: A survey. *ACM Comput. Surv.*, 55(8).

Watheq Ahmad Mansour, Salam Albatarni, Sohaila Eltanbouly, and Tamer Elsayed. 2024. Can large language models automatically score proficiency of written essays? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2777–2786, Torino, Italia. ELRA and ICCL.

Sandeep Mathias and Pushpak Bhattacharyya. 2018. ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Atsushi Mizumoto and Masaki Eguchi. 2023. Exploring the potential of using an ai language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2):100050.

Tomoya Mizumoto, Hiroki Ouchi, Yoriko Isobe, Paul Reisert, Ryo Nagata, Satoshi Sekine, and Kentaro Inui. 2019. Analytic score prediction and justification identification in automated short answer scoring. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 316–325, Florence, Italy. Association for Computational Linguistics.

Christoph Molnar. 2025. *Interpretable Machine Learning*, 3 edition.

Swapnil Parekh, Yaman Kumar Singla, Changyou Chen, Junyi Jessy Li, and Rajiv Ratn Shah. 2020. My teacher thinks the world is flat! interpreting automatic essay scoring mechanism. *Preprint*, arXiv:2012.13872.

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

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Felipe Urrutia, Cristian Buc, Roberto Araya, and Valentin Barriere. 2025. Unsupervised automatic short answer grading and essay scoring: A weakly supervised explainable approach. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 38–54, Vienna, Austria. Association for Computational Linguistics.

Bronwyn Woods, David Adamson, Shayne Miel, and Elijah Mayfield. 2017. Formative essay feedback using predictive scoring models. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 2071–2080, New York, NY, USA. Association for Computing Machinery.

Wenbo Xu, Rohana Mahmud, and Wai Lam Hoo. 2024. A systematic literature review: Are automated essay scoring systems competent in real-life education scenarios? *IEEE Access*, 12:77639–77657.

Ruosong Yang, Jiannong Cao, Zhiyuan Wen, Youzheng Wu, and Xiaodong He. 2020. Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1560–1569, Online. Association for Computational Linguistics.

Haneul Yoo, Jieun Han, So-Yeon Ahn, and Alice Oh. 2025. DREsS: Dataset for rubric-based essay scoring on EFL writing. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13439–13454, Vienna, Austria. Association for Computational Linguistics.

Lui Yoshida. 2025. Are the Reasoning Models Good at Automated Essay Scoring? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 8388–8394, Suzhou, China. Association for Computational Linguistics.

A Appendix: Complementary Figures

A.1 Hyperparameters

Parameter	Value
Batch Size	16
Max Length	512
Epochs	20
Learning Rate	0.005
Loss Function	MSELoss
Early Stopping Patience	5

Table 5: Training hyperparameters for both essay- and sentence-level models.

A.2 Normalisation of Pseudo-Scores

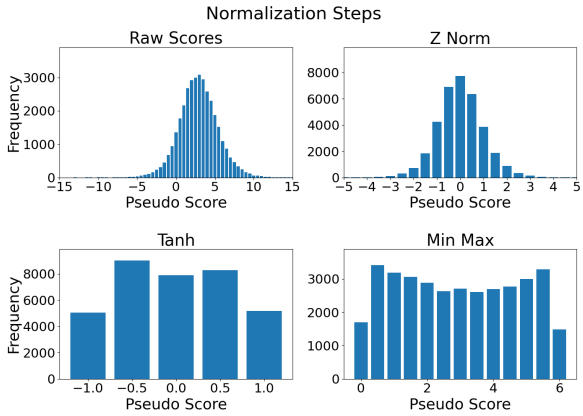

Figure 6: Distribution of sentence pseudo scores across the normalisation steps (z-norm, tanh, min-max) for one particular model.

B Appendix: Sentence-Level Performance

Dataset	Trait	Prompt	Essay	Sent	α	Abl.
MEWS	Content	AD	0.52	0.47	1	1
MEWS	Content	TE	0.45	0.41	1	1
MEWS	Content	TEAD	0.47	0.43	1	1
MEWS	Language	AD	0.70	0.71	1	1
MEWS	Language	TE	0.74	0.75	1	5
MEWS	Language	TEAD	0.72	0.74	1	1
ASAP	Content	1	0.69	0.40	1	1
ASAP	Content	2	0.67	0.61	1	1
ASAP	Content	3	0.60	0.49	1	1
ASAP	Content	4	0.68	0.62	1	5
ASAP	Content	5	0.73	0.70	1	1
ASAP	Content	6	0.69	0.69	1	1
ASAP	Content	7	0.67	0.53	1	1
ASAP	Content	8	0.65	0.24	1	5
ASAP	Language	3	0.57	0.48	1	1
ASAP	Language	4	0.63	0.54	1	5
ASAP	Language	5	0.70	0.54	1	1
ASAP	Language	6	0.61	0.56	1	1

Table 6: Performance (QWK) of essay- and sentence-level linear regression models.

Dataset	Trait	Prompt	Essay	Sent	α	Abl.
MEWS	Content	AD	0.33	0.29	20	1
MEWS	Content	TE	0.18	0.14	10	1
MEWS	Content	TEAD	0.30	0.23	1	3
MEWS	Language	AD	0.56	0.53	10	3
MEWS	Language	TE	0.41	0.58	20	1
MEWS	Language	TEAD	0.68	0.53	1	3
ASAP	Content	1	0.61	0.21	1	3
ASAP	Content	2	0.58	0.43	10	1
ASAP	Content	3	0.52	0.44	1	1
ASAP	Content	4	0.48	0.44	1	1
ASAP	Content	5	0.60	0.34	30	1
ASAP	Content	6	0.70	0.35	10	1
ASAP	Content	7	0.60	0.45	1	5
ASAP	Content	8	0.41	0.15	1	1
ASAP	Language	3	0.50	0.38	30	1
ASAP	Language	4	0.41	0.39	1	1
ASAP	Language	5	0.57	0.25	10	1
ASAP	Language	6	0.57	0.32	10	1

Table 7: Performance (QWK) of essay- and sentence-level LLM models.

Dataset	Trait	Prompt	Essay	Sent	α	Abl.
MEWS	Content	AD	0.44	0.35	1	1
MEWS	Content	TE	0.28	0.31	1	5
MEWS	Content	TEAD	0.33	0.45	1	1
MEWS	Language	AD	0.71	0.74	1	1
MEWS	Language	TE	0.46	0.81	1	5
MEWS	Language	TEAD	0.76	0.75	1	5
ASAP	Content	1	0.66	0.43	1	3
ASAP	Content	2	0.67	0.60	1	5
ASAP	Content	3	0.64	0.50	1	1
ASAP	Content	4	0.68	0.50	1	1
ASAP	Content	5	0.70	0.67	1	5
ASAP	Content	6	0.69	0.68	1	1
ASAP	Content	7	0.60	0.48	1	3
ASAP	Content	8	0.52	0.24	1	3
ASAP	Language	3	0.57	0.44	1	1
ASAP	Language	4	0.61	0.49	1	1
ASAP	Language	5	0.68	0.50	1	1
ASAP	Language	6	0.58	0.52	1	1

Table 8: Performance (QWK) of essay- and sentence-level Hybrid models.