ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

A Study on Intelligent Evaluation Methods for English Writing Proficiency Based on Semantic Feature Extraction

Pingping Liu[1,*]

[1] School of Education, Quanzhou Vocational and Technical University, Quanzhou 362000, Fujian, China

SUMMARY: *Addressing the issues of insufficient scoring granularity, weak cross-task stability, and unclear feedback evidence in university English writing instruction and intelligent grading scenarios, this paper proposes an intelligent evaluation method for English writing proficiency based on semantic feature extraction.First, a multi-source writing evaluation sample space is constructed around ASAP++, ICLE++, DREsS, and PERSUADE 1.0, and a unified mapping of composite scores, fine-grained traits, and rubric rules is established;Next, we establish a prompt-aware semantic feature extraction and multi-view joint scoring model, incorporating global semantic meaning of the prompt and essay, sentence-level coverage and local coherence, paragraph organizational relationships, and language support features into the scoring framework. By means of consistency constraint conditions, we have realized the combined forecast of the total score and five-dimensional meta-characteristics; In the end, experiments have been done by us under three evaluation schemes: inside-prompt, one-prompt-leave-out, and based-on-rubric. The results obtained by us indicate that our method attains a QWK of 0.862 and an MAE of 0.348 on the dataset ASAP++, attains a QWK of 0.821 for composite score and a trait-average QWK of 0.793 on the dataset ICLE++, and attains a Macro-F1 of 0.734 and a QWK of 0.796 on the dataset DREsS, hence all of these performances are better than the baseline methods. Experiments on low-resource conditions further prove that even when there is only twenty percent of training data used, the model still keeps a QWK of 0.713, hence this shows the model has good efficiency in utilizing samples. Case studies and error analysis indicate that this method can with reliability distinguish typical problems like inadequate topic answering, disconnected text development, and feeble language support, and hence has the possibility for being put into human-check working flows.*

KEYWORDS: *English writing proficiency assessment; automated essay scoring; semantic feature extraction; multi-task joint scoring; cross-task generalization*

1 Introduction

The assessment of English writing ability directly has influence on classroom feedback, period-based examinations and test judgment decisions. In university general English courses, writing practice platforms, and uniform language examinations, teachers must not only give a total score but also point out concrete problems such as content matching, text arrangement, word application, grammar correctness, and cohesion. Manual grading by human beings can take into consideration context, task purpose, and scoring standards, but in large-scale teaching situations, it frequently encounters practical limitations like long grading periods, non-uniform feedback

*ppsiobhan@126.com
https://doi.org/10.65102/is2026758
Published: 30 April, 2026

detail levels, and problems in unifying assessments among different classes. The core demand which comes from this is to build an intelligent assessment interface which can reliably support English writing examination, ensuring that marking results keep consistency with human judgements while offering enough detailed proof to support later teaching interventions.

This need manifests very concretely in real-world application scenarios. Classroom writing exercises emphasize timely feedback; teachers are concerned with whether students genuinely address the prompt, whether arguments are developed, whether paragraphs progress logically, and whether language issues are concentrated in specific areas. In examination settings, the emphasis shifts to consistent grading standards, transferability across different prompts, and manageable review costs.If an evaluation system can only output a single overall score, it is difficult for teachers to identify students' weaknesses in content, organization, or language based on that score; if the system is highly sensitive to variations in prompts, its practical value in open-ended writing exercises will rapidly diminish. Intelligent evaluation of English writing proficiency must address not only the issue of "scoring speed" but also the synergy among three dimensions: scoring granularity, cross-prompt stability, and the usability of feedback.

For solving these difficulties, the study of automatic essay score giving has slowly moved from early methods that pay attention to feature engineering and linear regression to deep representation learning, long text encoding, and fine-grained character assessment. Latest literature summaries show that the present research field is experiencing two big changes: first, models do not any more only pay attention to matching one single whole score but put bigger stress on separate evaluations on aspects including content, organization, vocabulary, grammar, and coherence; second, the research emphasis is now moving from closed-type assessments on same prompts to cross-prompt generalization, rule-bound assessment, and application stability [1, 2]. Under this background, score frameworks which combine multi-angle expressions—including semantic, topic, and language aspects—show higher explainability. The study about scoring that uses LLM for essays of English learners has also caused validity, reliability, and bias control to return to the front position of methodological discussions[3, 4].

The rise of cross-prompt and fine-grained evaluation has transformed the task definition of automated essay scoring.Many early methods performed well on a single open benchmark, but model performance declined significantly when the prompt changed, the genre shifted, or the rubric was adjusted. Li and Ng's systematic analysis of cross-prompt scoring noted that differences between prompts in task intent, genre constraints, and scoring boundaries make it difficult to directly transfer representations learned from the source prompt to the target prompt [5]. Accordingly, recent datasets such as ICLE++ and DREsS have shifted the focus from a single overall score to multidimensional trait- and rubric-based evaluation, bringing automatic essay scoring closer to how it is used in real classroom and EFL settings [6, 7]. This shift implies that models must address two questions simultaneously: whether they can maintain consistent scoring across different prompts, and whether they can decompose scoring results into competency dimensions that are understandable and traceable to teachers.

As a reply to these changes, the current researches have already started to deal with these difficulties from many different angles. Some researches promote cross-task generalization via syntactic amendment or prompt-generalized expression forms [8]; other researchers bring in ordered expert frameworks or rating rule limits to raise the stability of character scoring [9-11]; and some use large language models to produce scoring reasoning, with the goal of promoting the interpretability of multi-dimensional feedback [10]. In addition, graph-based methods have been used to construct the dependence relationships between traits, hence reducing local inconsistent problems in multi-dimensional scoring. [12]. These progressions show that automatic essay scoring has moved its emphasis from "if a score can be matched" to "if steady judgments can be set up on the foundation of scoring proof."

However, the nowadays methods still have three key shortcomings in the actual English writing evaluation situations. First, at the semantic representation level, there remains an overemphasis on compressed representations of the entire essay, with insufficient joint characterization of the prompt's semantic units, sentence response density, and paragraph development logic. For the students who study English and write essays, it has not been found that there is a fixed corresponding relation between the repetition of key words and the high-quality answer compositions; without the control of sentence-level semantic coverage and redundancy, models are easy to mistakenly regard the surface compliance to the prompt as substantial content. Second, though many models possess multiple scoring heads, they are not able to bring textual progression and local coherence evidence into a unified representation space, therefore making the organizational and coherence dimensions still easy to receive interference from text length, sentence length, and surface lexical complexity. Third, along with the large language models (LLMs) that are integrated into the scoring process, discussions which concern demographic bias, scoring uncertainty, and the replaceability of high-risk scenarios can no longer be avoided by people. Existing research suggests that models may amplify differences among certain marginalized groups, while a single numerical score is insufficient to support direct replacement of human decision-making in application scenarios [13, 14]. Therefore, intelligent evaluation of English writing proficiency requires a method capable of integrating prompt semantics, essay content, text organization, and linguistic support evidence into a single evaluation framework, while retaining a confidence score interface at the output stage to provide a basis for human review.

Based on the above considerations, this paper narrows the research problem to the following: how to construct an intelligent evaluation method—centered on semantic feature extraction, balancing overall scores with trait-based scoring, and providing confidence metrics for deployment—across diverse prompts, fine-grained dimensions, and English learner writing scenarios.To address this problem, this paper proposes a prompt-aware semantic feature extraction and multi-view joint scoring model. Based on a global semantic encoding of the prompt and the essay, this model explicitly captures semantic coverage, adjacent coherence, and content redundancy at the sentence level; supplements text organization information at the paragraph level; and introduces lexical, syntactic, and grammatical support features at the linguistic level, thereby converging evidence from different sources into a unified scoring space.

The present work has three main contributions. First, we put forward a prompt-knowing semantic feature picking method for English writing evaluation, integrating whole-text semantic meanings, sentence-layer matching, text arrangement, and language support features into a single scoring frame system. Second, we establish a multi-source English writing assessment data set arrangement plan, realizing score standardization and five-dimensional meta-character matching among ASAP++, ICLE++, and DREsS, therefore we use PERSUADE 1.0 to strengthen discourse information construction. Third, we put forward consistency constraint items between the total score and scores on each individual dimension, hence we keep a human-machine cooperate checking interface for samples with low confidence, therefore we promote the model's usability in cross-scene English writing evaluation situations.

2 Methods

2.1 Multi-source Corpus Organization and Sample Construction

After defining the problem, this paper first addresses the issue of sample organization. No single publicly available dataset exists for English writing assessment that simultaneously covers overall scores, fine-grained traits, rubric rules, and cross-task transfer; therefore, the method

design cannot rely on a single dataset.To ensure the model can learn stable overall score boundaries while also being exposed to finer-grained scoring dimensions, this paper adopts an organizational approach of "main task corpus + fine-grained corpus + rubric corpus + text-based auxiliary corpus," incorporating ASAP++, ICLE++, DREsS, and PERSUADE 1.0 into a unified training and validation framework [15, 16].The purpose of this organizational approach is clear: ASAP++ provides a comprehensive score benchmark for traditional automatic essay scoring; ICLE++ provides fine-grained trait supervision for English learner essays; DREsS provides a rubric-based EFL writing evaluation scenario; and PERSUADE 1.0 provides auxiliary information on paragraph roles and text components.

To clarify the training targets, label sources, and the roles of each corpus in the task, we first summarize the composition and purposes of the corpora used, as shown in Table 1.

Table 1: Corpus Component, Label Rules, and Utilization in This Research

Corpus	Sample Size	Number of Topics/Themes	Label Granularity	Purpose of This Article
ASAP++	12,978	8	Overall Score + Attribute Score	Within-prompt benchmark comparison, score normalization base
ICLE++	1,008	10	Overall Score + 10 Fine-grained Traits	Cross-topic trait evaluation in EFL writing
DREsS	48,900	Multiple Topics/Multiple Rubrics	Rubric-based Scoring	Generalization testing and low-resource stability analysis
PERSUADE 1.0	25,996	15	Discourse Element Annotation	Paragraph roles and discourse structure assisted supervision

In Table 1, four kinds of corpus have obvious distinctions on sample quantity, amount of prompts, and scoring fineness degree. ASAP++ has as its goal the construction of a complete marking framework which matches current studies; ICLE++ furthermore expands its attention to the fine-grained character level of EFL writing; DREsS puts emphasis on scoring stability in rubric-based and cross-task situations; PERSUADE 1.0 does not directly make contribution to the main results statistics, but it offers auxiliary supervision for the modeling of paragraph role identification and discourse progression. This division of data sets lets the model at the same time get overall score judgments, scores for each specific dimension, and information of discourse structure, hence it avoids wrong understanding of local strengths on one single public benchmark as overall abilities.

Because scoring standards and feature names have differences among different text databases, this paper at the beginning carries out unified pretreatment, after that carries out label mapping. The preprocessing step eliminates empty text, disordered text, samples that have lacking scores, and extremely short invalid articles; carries out standardization work on character coding, punctuation marks, and clause dividing lines for prompt texts; and it keeps the original paragraph dividing for essays meanwhile it eliminates samples that only have the prompt or have no substantial content. As for essays that have abnormal lengths, this study does not carry out the action of deleting them directly. As an alternative, when their length has been finished calculating, they are made comparison with the quartile scope of samples that belong to the identical prompt. If the content density of these things is extremely low, and their length is more than two times the upper limit of the prompt group, therefore they are marked as

abnormal and hence are removed. This method is adopted because English learner writings contain a great many of boundary samples that are either "long but repeated" or "short but full"; Only depending on length critical values would bring about the wrong deleting of effective samples.

After the completion of cleaning work, this paper carries out normalization processing and mapping operation on the scoring labels. Composite total points are first scaled to the [0,1] range inside each individual question, then brought back to the original effective score group when the testing stage is processed, hence the comparability of scores between different questions can be guaranteed. With respect to traits, this paper uniformly maps the original labels onto five meta-traits: Content, Organization, Vocabulary, Grammar, Coherence. To speak specifically, within ICLE++, items that are related to stance expression, task completion, and argument support are all merged into Content; projects which are connected to paragraph arrangement, connection between paragraphs, and whole structure are put together into Organization or Coherence; The scope of vocabulary and the accuracy of word selection are put into the category of Vocabulary; and the correctness of grammar and the control of syntax are put into the category of Grammar. DREsS rubric labels are got together according to the main evaluation dimensions which are written out in their scoring rules. If a certain language corpus does not have the corresponding characteristic, that dimension is covered in the training process and is excluded from the loss backpropagation computation. By means of this method, the comprehensive score and scores of each dimension may be combined into the identical training space, without the need to make the assumption that the scoring rules of different text corpora are completely isomorphic.

The construction of samples also need to solve the problem of alignment between prompt texts and essay texts. This paper at first carries out semantic unit cutting on the prompts, decomposing them into four kinds of smallest units: task movements, content limits, stance demands, and restrictions. As an example, for argumentative essay topic prompts, the model records requirements including "whether one stance is taken," "whether reasons are given," and "whether the essay deals with the appointed topic"; As for explanation tasks or tasks that need giving answers, the model keeps units including descriptions of objects, explanations of cause and effect, and restrictions of tasks. This step cannot straight produce scoring labels, but it decides the calculation granularity of following sentence-level semantic coverage, hence it hence influences the classification boundaries for the Content and Coherence dimensions.

With respect to the division agreement, this present article uses three assessing methods to carry out parallel checking. For ASAP++, we use five-fold inside-prompt cross-checking in order to assess the basic scoring capability of the model upon traditional standard test datasets; For ICLE++, we use a "leave-one-prompt-out" method, hence we rotate the excluding of prompts among 10 questions, therefore to evaluate the stability of trait scoring that is across prompts; About DREsS, we arrange the training, validation, and test sets in accordance with its standard division, and moreover build low-resource sub-sets with training proportions spanning from 20% to 100% to carry out analysis on the model's robustness in situations where labeling is not enough. The extraction work for low-resource small sets was not simple random pick, but was instead layer-divided sampling that based on question, score grade and length percentage, therefore it can hold a steady difficulty structure among all different training proportion ratios. The auxiliary corpus PERSUADE 1.0 was only utilized for training the paragraph role recognizer and the discourse transition prompt generator, and it did not directly take part in the scoring of the test set.

In terms of training input organization, this paper treats each essay as a four-layer object: "prompt + essay + sentence sequence + paragraph sequence." The prompt-essay joint input is used for global semantic encoding; the sentence sequence is used to calculate prompt response

density, local coherence, and content redundancy; and the paragraph sequence is used to construct a discourse organization graph.The purpose of this approach is to enable evidence at different granularities to share the same essay while each assuming distinct evaluation responsibilities: global representations are responsible for identifying overall intent, sentence-level representations for defining local response boundaries, paragraph representations for organizational progression, and linguistic metrics for correcting samples that are "semantically relevant but linguistically deficient."

For explaining how multi-source language materials change from different marking systems to one combined sample space, this paper gives a arranged introduction of the data arrangement and sample building procedure, which is displayed in Figure 1.

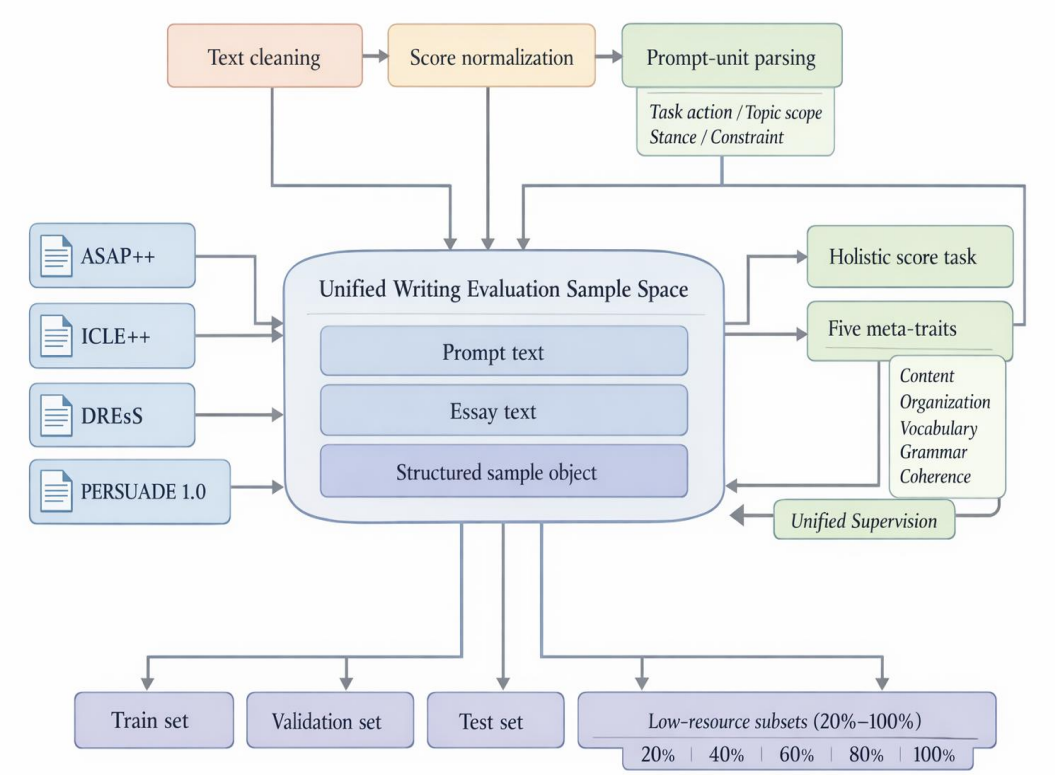

Figure 1: Diagram of Multi-source English Writing Corpus Organization and Sample Construction.

In Figure 1, SAP++ acts as the overall scoring reference standard, ICLE++ offers fine-grained feature marks, DREsS is utilized for rubric-based EFL generalization verification, and PERSUADE 1.0 works only as a text-level assisting supervisor. Because there exist big differences between different corpora in the aspects of sample size, quantity of prompts and label granularity, therefore it is necessary that we first carry out score normalization and meta-trait mapping before we enter the unified training space. If not do this, the overall scores, characteristic scores, and text layer supervision can not be carried out comparison on a unified training dimension.

2.2 Prompt-Aware Semantic Feature Extraction and Multi-View Scoring Model

In the design of the scoring model, this paper focuses on three core issues: whether the essay genuinely addresses the prompt, whether there is effective progression between paragraphs, and

whether the linguistic form supports the expression of its content. Addressing these three issues, the model consists of a global semantic branch, a sentence-level alignment branch, a discourse organization branch, and a linguistic support branch, ultimately performing joint prediction of the composite score and trait scores at a unified scoring layer [17, 18].The global semantic branch takes the prompt text and the full essay as input to obtain an overall representation of the entire essay under the current task conditions. This branch is primarily responsible for identifying the prompt's intent, stance, and topic relevance, providing the foundational semantic representation for the comprehensive score.Relying solely on the full-text representation can easily lead to misclassifying "seemingly relevant" content as "task-compliant responses." Therefore, this paper further incorporates a sentence-level alignment branch to explicitly compute the correspondence between each sentence and the semantic units of the prompt. The core variables of the sentence-level alignment branch are shown in Equation (1).

$$s_i = \max_j \cos\left(e_i, p_j\right) \tag{1}$$

In the formula, s_i represents the semantic coverage score of the i th sentence relative to the prompt, e_i represents the semantic vector of that sentence, and p_j represents the vector of the j th semantic unit in the prompt. A higher score indicates a stronger alignment between the current sentence and the key requirements of the prompt. Based on the sequence distribution of s_i, the model can distinguish three common scenarios: consistent and stable response to the prompt, alignment in the first part followed by deviation in the latter part, and excessive repetition of prompt keywords without substantial elaboration[19-22]. Compared to overall similarity, this local coverage metric is better suited for handling borderline samples in English learners' essays where the text "surface-level aligns with the prompt but lacks substantive content."

Beside the sentence-level coverage, this paper likewise carries out calculation of the semantic closeness between neighboring sentences and the maximal overlapping between the current sentence and the sentence that comes before it. The first one manifests local coherence, hence the second one discovers repeatable arguments, overmuch pileup of examples, and not enough advance of information. These measurement indexes are not used one by one as final scores but are put together to make a sentence-level semantic description vector. This vector mainly undertakes the two dimensions of Content and Coherence, and it also can provide partial proof to make the composite score more complete.

The discourse arrangement branch further regards paragraphs as writing units that have clear functions, not just text blocks that are arranged in order one after another. This present paper firstly carries out encoding work for every individual paragraph, after that it next builds up a graph of paragraph relations through the combination of paragraph position, semantic similar degree and discourse marking labels. In the graph, the nodes stand for paragraphs, hence the edges indicate the progressive connections which are between them. For the purpose of promoting the stability of organizational judgments, this paper utilizes a paragraph role classifier which is trained through PERSUADE 1.0 to offer weakly supervised restrictions for functional positions, for example, introduction, claim, evidence, conclusion. Therefore, the model places emphasis not merely on whether a paragraph "exists", but also on "what function it undertakes" and "whether it pushes forward the core mission."

The language support branch extracts quantitative indicators such as vocabulary variety, average clause thickness, dependency length, grammar mistake density, percentage of discourse linking words, and punctuation abnormality rate from the complete text. This branch does not hold the dominant position in scoring direction but undertakes the correction work for potential overestimation results which are produced by the semantic branch. For instance, certain written

works may display high topic-related connection but include lasting groups of grammar mistakes or show obvious shortcomings in sentence structure management. Under these circumstances, the scores of Grammar and Vocabulary should not be raised by the overall similarity of meaning. On the opposite side, if the language is comparatively smooth, but the answer to the task is not enough, the scores of Content and Organization hence should be kept at low levels.

After the four feature categories enter the unified scoring layer, the model outputs five meta-trait scores and a composite score. The unified scoring vector is defined as shown in Equation (2).

$$\hat{y} = W\,[g; u; d; l] + b \tag{2}$$

where $\hat{y}$ denotes the scoring output vector, g represents global semantic features, u represents sentence-level alignment and local coherence features, d represents text organization features,l represents language support features, $[\cdot]$ denotes vector concatenation, and$W$ and$b$ are learnable parameters. This output vector contains five trait scores as well as a predicted composite score. This design aims to ensure that the composite score is no longer generated in isolation but remains consistent with the local assessments across each dimension.To reduce the logical discrepancy between the composite score and the trait scores, this paper introduces a joint loss function during training, as shown in Equation (3).

$$\mathcal{L} = \mathcal{L}_h + \lambda_1 \mathcal{L}_t + \lambda_2 \mathcal{L}_c \tag{2}$$

In the equation, $\mathcal{L}_h$ represents the overall score regression loss, $\mathcal{L}_t$ represents the trait score regression loss, $\mathcal{L}_c$ represents the consistency constraint term, and λ_1 and λ_2 are loss weights. The consistency constraint is used to suppress output imbalances such as "generally low scores across dimensions but an abnormally high overall score" or "an abnormally inflated score for a single trait." For samples with missing trait labels, only the overall score loss and the trait losses for observable dimensions are calculated.

Besides the scores themselves, this article also records the variance of many dropout inferences on the output side, and combines it with the temperature-scaled probability which the fusion layer represents, to constitute a confidence score. When one sample satisfies two conditions together, which are "high prediction variance" and "low calibrated confidence," hence it will enter the manual review channel [23-25]. This design lets the model not merely give a score but also evaluate the stability of this score, hence it provides a risk control interface that can be put into practice for following arrangement and application. After we have finished the design work of the feature and the scoring structures, the internal information interactions and constraint relationships that are inside the model can be represented through Figure 2.

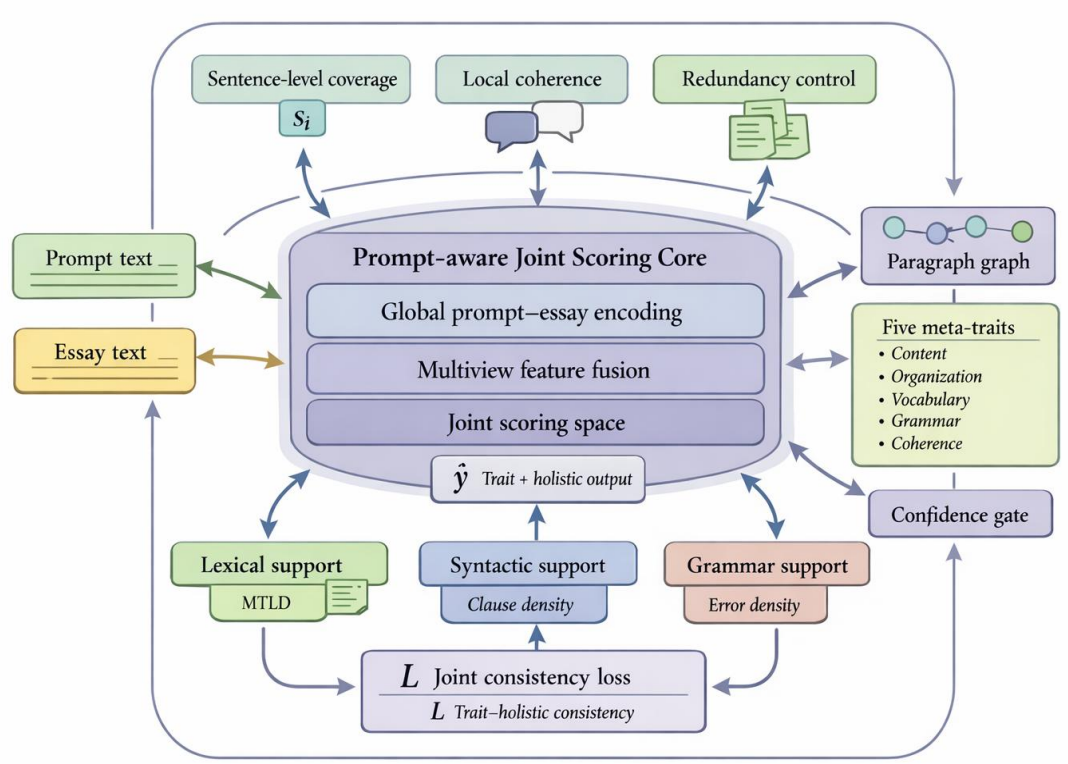

Figure 2: Prompt-aware semantic feature extraction and multi-view scoring diagram.

In Figure 2, global encoding is responsible for extracting the overall semantics of the prompt and essay; sentence-level alignment identifies the density of prompt responses, local coherence, and content redundancy; the paragraph relationship graph characterizes the progression of the text; and language support features are used to correct borderline samples that are "semantically relevant but linguistically deficient."After fusion, each branch simultaneously outputs trait scores, composite scores, and confidence levels, enabling the model to provide feedback across multiple dimensions while preserving a discrimination interface for subsequent manual review.

2.3 Training Strategy, Comparison Methods, and Evaluation Protocol

Regarding the training strategy, the present paper utilizes the AdamW optimizer, having a maximal input length of 1024, a batch size of 8, an initial learning rate of 2×10^{-5}, and at most 12 training epochs. The training process is brought to an early stop in the case that the validation measurement indicators do not have any enhancement for three continuous training rounds. In order to alleviate random undulations, all experiments have been done five times repeatedly, and the average outcomes are put forward. The main encoder makes use of a lower learning rate, hence the scoring layer and discourse organization layer make use of a higher learning rate to control the update speed between pre-trained semantic expression representations and task-special parameters.

To discrete grading data, a continuous regression type is all used in the training stage, and the outcomes are mapped again to lawful grading grades in the testing stage. This method alleviates gradient shake movements which are brought by discrete boundary limits, and it meanwhile retains the explainable property of the rating grade levels. For data which have missing trait labels, the model skips the loss backpropagation of the corresponding dimension through a masking mechanism, hence it ensures that samples from different sources can join the training inside the same framework.

The comparison methods cover three representative approaches. The first category consists of traditional cross-item scoring baselines, which serve as low-complexity reference points for comparison. The second category comprises scoring methods based on pre-trained language models, used to observe the upper limit of full-text semantic encoding.The third category comprises recent multi-task, trait-aware, and prompt-aware models, used to identify the sources of improvement for our method in cross-question and sub-dimension scoring tasks. All baselines are reproduced under a unified training interface, while maintaining the input organization and hyperparameter ranges recommended by the original methods as much as possible.

With respect to evaluation metrics, both the synthesized total score and each individual trait score are all reported through QWK, MAE, and Pearson correlation coefficients. For the scenarios that are based on rubrics, Macro-F1 is additionally reported for reflecting the balanced performance among all the different grading levels. QWK is utilized for measuring the consistency with human scoring, MAE reflects the mean absolute error value, the Pearson correlation coefficient reflects continuous tendency change, and Macro-F1 is utilized for evaluating classification effect under the situation of imbalanced grade distribution. Significance examinations utilize paired t-tests, with a threshold being settled at $p < 0.05$.

For the purpose of making certain that the obtained outcomes exceed the scope of pure score comparisons, this paper has additionally designed an error analysis and examination working procedure. In the test gathering, if the prediction mistake for the combination score surpasses one effective evaluation grade, the specimen is marked as a serious wrong classification. All samples with serious wrong classification, together with a same quantity of randomly chosen ordinary samples, were undergone manual checking by people. The review dimensions contain inadequate reaction to the topic requirement, disconnected paragraph arrangement, gathered grammar mistakes, repeated key words, and controversies concerning grading scope boundaries. Through this protocol, the results got from the model not only can be utilized for the primary comparisons of performance, but also can be utilized to analyze the concrete regions of risk inside the actual teaching situations. For preventing wrong understanding of outcomes which comes from differences of experiment parameters, this paper gives out unified restrictions to the processes of training, comparison and evaluation, as what is shown in Figure 3.

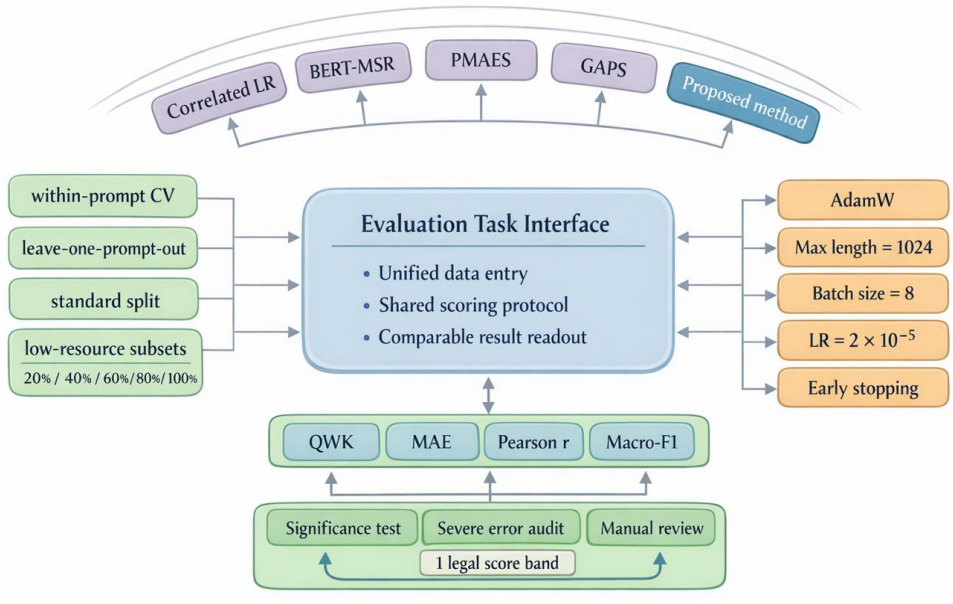

Figure 3: Diagram of experimental setup, comparison strategies, and evaluation protocols.

In Figure 3, the "inside-prompt," "put-one-prompt-out," and DREsS standard classification methods together compose three verification situations. Both basic baseline methods and our put-forward method all obtain QWK, MAE, Pearson r, and Macro-F1 under the identical evaluation platform. Samples which have low confidence are then sent into a manual check flow, therefore ensuring that performance comparison, error finding, and deployment choices are completed inside one single protocol frame, rather than being spread over mismatched experiment arrangements..

3 Results and Discussion

3.1 Overall Performance and Cross-Task Generalization Results

After we have finished the work of data arrangement and model training, this research carries out an examination on whether the method that we have put forward can keep a stable superiority in both the comprehensive scoring and the cross-task evaluation. If one model has good performance on only one metric but has obvious failure in the writing of English learners or the situations based on rubrics, hence the design that is based on semantic feature extraction does not have enough support. According to this evaluation, this research firstly carries out a comparison of whole performance, and then makes an investigation of changes under low-resource situations and among particular characteristics. After we finish the unified training and testing, we first carry out comparison on the whole performance of different methods on three data metrics, which is displayed in Table 2.

Table 2: Comparison of Overall Performance Across Multiple Datasets

Method	ASAP++ QWK	ASAP++ MAE	ICLE++ Holistic QWK	ICLE++ Trait Mean QWK	DREsS Macro-F1	DREsS QWK
Correlated LR	0.781	0.456	0.703	0.668	0.621	0.684
BERT-MSR	0.812	0.411	0.742	0.709	0.658	0.721
PMAES	0.829	0.392	0.761	0.734	0.676	0.739
GAPS	0.838	0.381	0.774	0.748	0.689	0.751
TRATES	0.844	0.376	0.791	0.769	0.706	0.768
Proposed Method	0.862	0.348	0.821	0.793	0.734	0.796

In Table 2, the method put forward by us has obtained the best outcomes in all three situations: ASAP++, ICLE++, and DREsS. When we make a comparison with the strongest baseline method, TRATES, our method has a 0.018 improvement on QWK and a 0.028 reduction on MAE over the dataset of ASAP++; on the data set of ICLE++, the overall QWK score has a promotion of 0.030, therefore the average trait QWK hence got an improvement of 0.024; On the dataset DREsS, both the Macro-F1 index and the QWK index have the increase of 0.028. These outcome show that the strong points of our method do not only restrict on traditional composite marks, hence, they are more obviously embodied on English learner writing pieces and rubric-related cross-task assessment. This is on account of that prompt-aware semantic feature extraction lets the model go beyond dependence on compressed expressions of the whole essay, hence it incorporates topic reply, paragraph-level development, and language support into the scoring standards. Therefore, hence, the model can keep boundary stability in a more effective way in EFL situations which have the features of big changes in

essay topics and larger language change degrees. For the deeper examination of how labeling scale influences model robustness, we have made comparison of cross-essay performance in the condition of different training proportions, which is displayed in Figure 4.

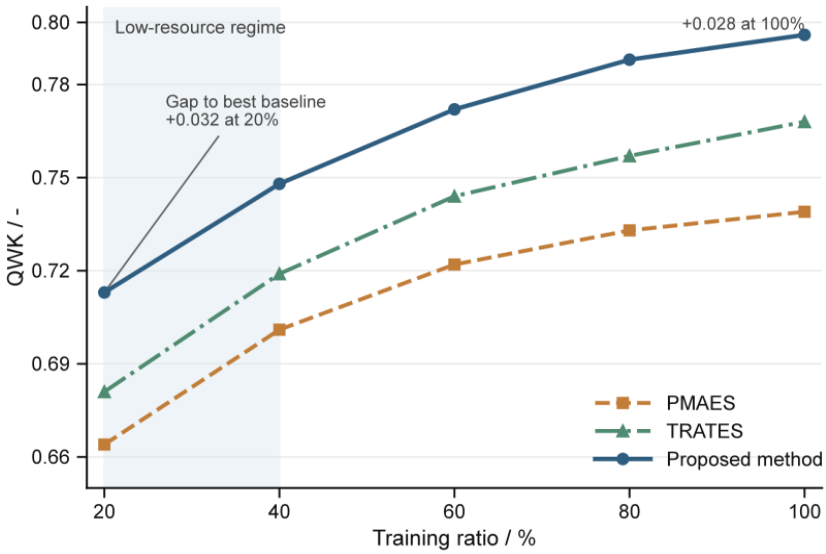

Figure 4: QWK variation curves of models under different training ratios.

In Figure 4, as the training ratio increases from 20% to 100%, the QWK of all models continues to rise, but our method consistently remains on the highest curve.Under low-resource conditions (20%), our method achieves a QWK of 0.713, outperforming PMAES (0.664) and TRATES (0.681); when the training ratio is increased to 100%, these values rise to 0.796, 0.739, and 0.768, respectively.These results indicate that prompt-sentence alignment and paragraph relationship modeling are more effective when data is scarce, as they provide the model with finer-grained attribution cues than full-text compressed representations, enabling the model to maintain a basic understanding of task-relevance and textual progression even with limited annotations.At the same time, the slope of the increase in our method from 20% to 60% is significantly steeper than that of PMAES, indicating that our framework is more efficient at absorbing newly added annotations. Beyond the overall scores, the results by dimension better reflect whether the model has truly captured the structure of writing ability, as shown in Figure 5.

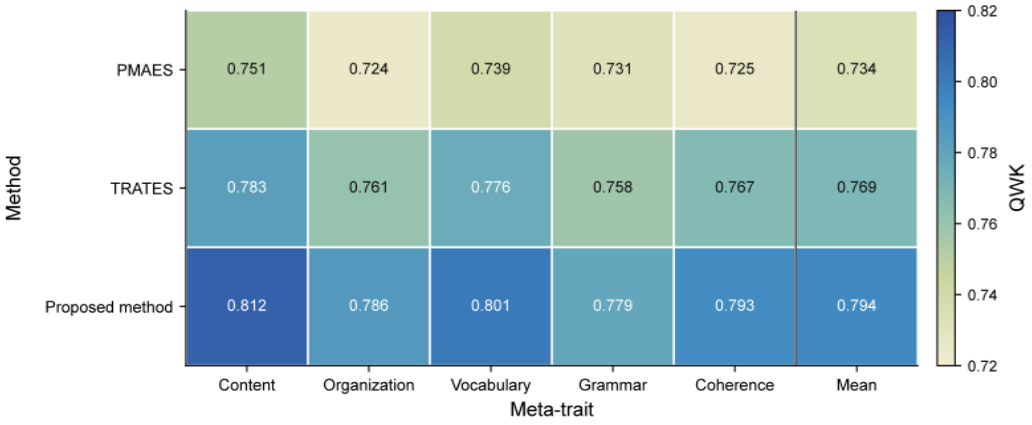

Figure 5: QWK heatmap of ICLE++'s five-dimensional meta-traits.

In Figure 5, the method proposed by us holds the topmost QWK in all five meta-traits—Content, Organization, Vocabulary, Grammar, and Coherence—with the score values being 0.812, 0.786, 0.801, 0.779, and 0.793, respectively, all of which are higher than TRATES' score values of 0.783, 0.761, 0.776, 0.758, and 0.767. Among all these aspects, the promotions of Content and Coherence are the most obvious, which attain 0.029 and 0.026, respectively each. The previous one points out that sentence-layer semantic coverage can more well distinguish between "relation to the question request" and "true task-oriented replies"; this latter viewpoint holds that paragraph relation graphs and partial consistency characteristics can supply extra information for discourse-layer evaluation. Word stock and structural arrangement also display steady increments, hence this shows that semantic matching does not shrink the feature space of language assistance features. The increment of Grammar is comparatively tiny, which is connected with the highly localized and fine-grained character of grammatical mistakes in English learners' writing, hence it suggests that it is still necessary to introduce finer-grained error type representations in future work.

3.2 Module Ablation, Efficiency, and Error Source Analysis

The whole outcomes prove the usefulness of our method; however, deeper analysis is required to make certain which modules have contribution to the promotion of performance, whether the calculation cost can be controlled, and in which places errors are mainly gathered. For this purpose, this section carries out examination on these aspects via module ablation, inference cost, and error samples. Although the whole outcomes have already proved the merits of our method, further analysis is needed to find out which modules push forward the performance promotion and whether the calculation cost is controllable, like what is displayed in Table 3.

Table 3: Module Ablation and Efficiency Statistics

(a) Module Ablation Results

Setting	ASAP++ QWK	ICLE++ Trait Mean QWK	DREsS QWK
Remove Prompt Alignment Module	0.848	0.779	0.783
Remove Discourse Graph Module	0.851	0.781	0.786
Remove Language Support Features	0.854	0.784	0.789
Change to Single-Task Holistic Scoring	0.845	0.772	0.779
Complete Model	0.862	0.793	0.796

(b) Number of Parameters and Inference Cost

Method	Parameters / M	Delay / ms per essay	VRAM / GB	QWK
BERT-MSR	110	38	7.4	0.812
PMAES	125	47	8.3	0.829
TRATES	210	112	14.8	0.844
Proposed Method	156	61	9.6	0.862

In Table 3(a), the performance advantage of the full model does not stem from the independent contribution of a single module, but rather from the combined effect of multiple types of evidence. If the prompt alignment module is removed, the QWK on DREsS drops from 0.796 to 0.783—the largest decrease—indicating that, under cross-prompt conditions, the first capability to be lost is the fine-grained assessment of how well the response addresses the prompt;After removing the text map module, the average QWK for the trait on ICLE++ drops to 0.781, indicating that paragraph progression relationships play an irreplaceable role in assessing Organization and Coherence; although the decline is relatively small when removing

language support features, it still results in a consistent drop across all three metrics, suggesting that language features serve a boundary correction function within this framework.If the model is modified to predict only the composite score, the average QWK of ICLE++ traits drops to 0.772, indicating that joint training indeed improves the consistency between the total score and the scores across individual dimensions.In Table 3(b), our method achieves a QWK of 0.862 with 156M parameters and a latency of 61 ms/essay, outperforming TRATES, which has 210M parameters and a latency of 112 ms/essay.Compared to TRATES, our method reduces the number of parameters by 54M and shortens the inference latency per essay by 45.5%, while still improving QWK by 0.018. This implies that the performance improvement does not rely solely on scaling up the model, but rather stems from more effective feature organization and scoring coupling methods. For educational platforms and university-level grading systems, this balance between accuracy and computational cost is more practical than simply pursuing the highest possible score.To further examine the consistency between model outputs and human scores, we plotted the correspondence between predicted scores and human scores, as shown in Figure 6.

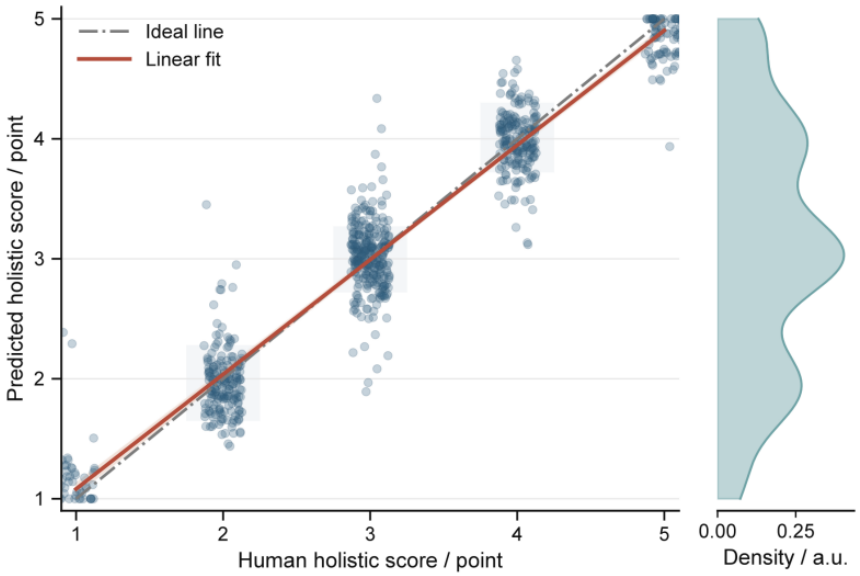

Figure 6: Scatter plot of consistency between predicted scores and human grades.

In Figure 6, most samples are distributed along the diagonal, indicating that the model is already capable of consistently reproducing the grade order assigned by human graders. The scatter is most concentrated in the middle score range, suggesting that the model is most stable in grading essays of average quality; the dispersion is higher at the low- and high-score ends, indicating that extreme samples still rely more heavily on finer-grained linguistic errors and argumentative quality cues.Further statistical analysis reveals that samples with an absolute error exceeding one grading level account for 4.7%, with the majority concentrated in essays categorized as "superficially on-topic but lacking in argumentation" and "relatively long but highly repetitive." The slope of the main scatter plot cluster relative to the fitted line is close to 1, which further demonstrates that the improvement achieved by our method is not driven by a small number of outliers, but rather reduces prediction bias across most intervals.Examining only the latency of a single model or the optimal value of a single metric is insufficient to address trade-offs in deployment. To place scoring accuracy and deployment costs within the same evaluation framework, this paper jointly plots the QWK of different methods against the inference latency per essay, as shown in Figure 7.

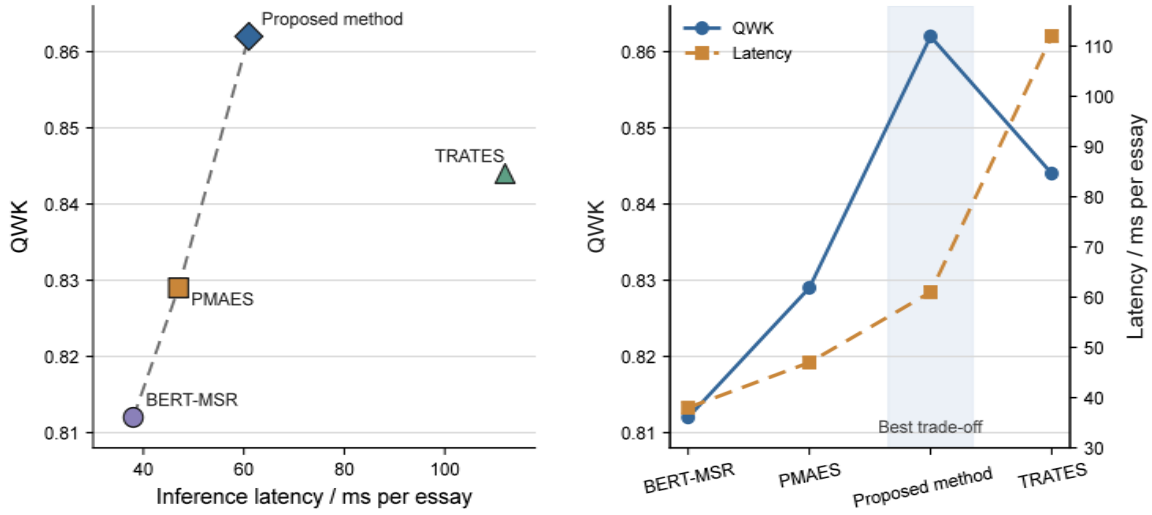

Figure 7: Accuracy-Latency Pareto Front for Different Methods.

In Figure 7, BERT-MSR, PMAES, and our method form the frontier of the current comparison set, while TRATES lies to the right of the frontier and is dominated by our method in terms of both QWK and latency.Specifically, our method achieves a QWK of 0.862 with a latency of 61 ms/essay, reducing latency by 51 ms/essay compared to TRATES while increasing QWK by 0.018; compared to PMAES, our method incurs an additional inference cost of 14 ms/essay in exchange for a 0.033 QWK improvement.Figure 7 does not simply represent "placing several model points in the same coordinate system," but rather illustrates how much accuracy gain can be obtained per unit of latency cost, as well as which models lie on the acceptable trade-off frontier. This result indicates that the benefits of our method primarily stem from more effective feature organization, rather than simply stacking models with larger parameters.

Regarding error sources, we manually reviewed the 100 essays with the largest absolute errors in the test set.The results showed that 34 essays fell into the category of "content on-topic but insufficiently argued," 29 essays fell into the category of "disjointed paragraph transitions leading to over- or underestimation of organizational coherence," 21 essays fell into the category of "keyword repetition leading to artificially inflated semantic relevance," and the remaining 16 essays were near the scoring boundary, reflecting discrepancies in the human scoring itself.This indicates that the samples the current model struggles most with are those that are superficially semantically relevant but lack sufficient argumentative depth. In other words, while semantic feature extraction has significantly improved the assessment of topic relevance, addressing higher-level quality issues such as "sufficiency of argumentation" requires stronger representations of argumentative structure and evidence chain modeling.

3.3 Typical Cases and Implications for Deployment

Overall statistics can demonstrate the model's effectiveness, but they cannot directly answer whether the model truly identifies the key aspects of writing quality. To determine whether methods based on semantic feature extraction are interpretable, we must return to the essays themselves. Based on this, this paper selects one high-scoring essay and one low-scoring essay to compare their topic semantic coverage, paragraph development methods, and trait scoring results, and further discusses the boundaries of human-machine collaboration during deployment.To observe whether the model truly bases its judgments on evidence of writing quality, this paper selects two representative essays for case analysis, as shown in Figure 8.

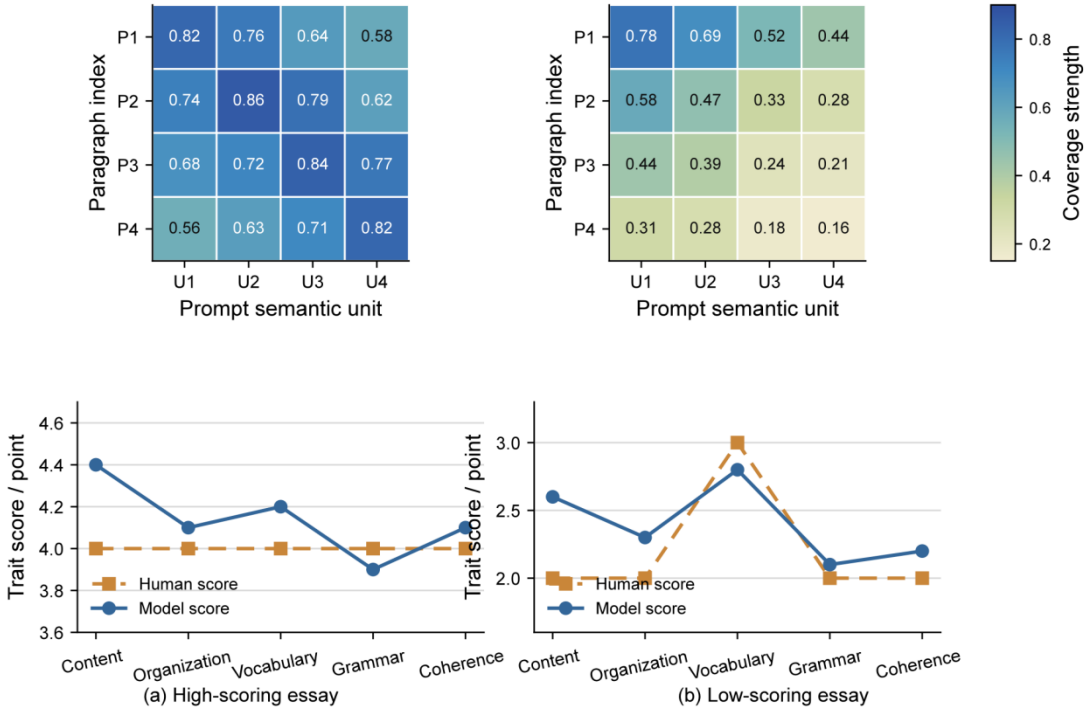

Figure 8: Case study of semantic coverage and trait scores for typical essays.

Figure 8(a) is corresponding to the high-score sample. This treatise puts forward its standpoint very clearly in the beginning paragraph, carries out the elaboration of reasons and examples around the requirements of the topic in the two paragraphs in the middle, and therefore winds up the discussion by pulling the argument back to connect with the given topic. The semantic unit and paragraph covering heat maps for this essay display a comparatively even distribution, especially in the argument paragraphs, where the covering is the strongest for the three semantic units: "position statement," "reason expansion," and "reaction to task limits." The model gave out marks of 4.4, 4.1, 4.2, 3.9, and 4.1 for Content, Organization, Vocabulary, Grammar, and Coherence, respectively, with a total mark of 4.2; the scores given by human beings were 4, 4, 4, 4, and 4, in order, hence they display a trend that on the whole is consistent. Here we need to point out that this model did not randomly expand the total score only because this sample possessed a big vocabulary; on the contrary, it still kept comparatively restrained upon the Grammar dimension, which shows that language support functions have played a role in adjusting the whole evaluation.

Figure 8(b) is corresponding to one sample that has low score. The first paragraph of this essay has completed the restatement of the topic in a successful way, but the second and the third paragraphs have begun to depart from the core task. Many sentences repeat the same opinion, there exists a shortage of effective development between paragraphs, and the latter half has a comparatively high gathering of subject-verb agreement and tense mistakes. When we compare with high-score samples, the coverage of its semantic units in the prompt is more concentrated on the keywords themselves, not on the core semantics of the task requirements, hence the heatmap shows a decreasing trend of "high in the beginning, low in the later part". The model gave out marks of 2.6, 2.3, 2.8, 2.1, and 2.2 to Content, Organization, Vocabulary, Grammar, and Coherence, separately, hence it obtained an overall mark of 2.4; The artificial evaluation scores are 2, 2, 3, 2, 2, in that order. Although there existed small differences in the scoring results, two kinds of evaluations all pointed out that the main problems are organizational structure, grammatical correctness and coherence. This contrast research shows

that the scoring standards of the method which this paper introduces are able to assess question answer, paragraph expansion, and language problems inside one unified frame, hence not only concentrating on surface word repetition.

Beyond case analysis, deployment scenarios are more concerned with determining when the model should automatically output results and when it should be handed over for human review. Based on threshold calibration using the validation set, this paper marks samples with a prediction variance higher than 0.62 as low-confidence samples and transfers them to human review.Under this configuration, the system can directly assign automatic scores to 78.4% of the essays, while the remaining 21.6% of samples enter the manual review process; if a "comprehensive score deviation exceeding one grade level" is considered a severe misjudgment, the severe misjudgment rate can be reduced from 6.8% in fully automatic mode to 2.1%.These results indicate that, in the context of English writing assessment, a more reasonable deployment approach involves using a combined output of semantic features and confidence scores to identify borderline samples, which are then subject to a secondary judgment by teachers or graders. This approach preserves the efficiency advantages of automated scoring while reducing the cost of misjudgment for high-risk samples.

Furthermore, another issue in deployment is fairness. Recent studies indicate that models may amplify existing scoring disparities across certain demographic attributes or linguistic backgrounds. Based on this assessment, our method adheres to two principles in its deployment recommendations: First, identity attributes such as gender, race, family background, or native language should not be used as inputs;Second, after the system goes live, QWK, the severe misjudgment rate, and the proportion of cases referred to human review should be monitored by question type, grade level, and linguistic background. Only by integrating performance monitoring and human review interfaces into a single workflow can intelligent evaluation of English writing proficiency transition from laboratory results to an application form acceptable for both peer review and teaching.

4 Conclusion

Addressing the practical challenges of rapidly changing prompts, multiple scoring dimensions, and high costs of human feedback in English writing assessment, this paper proposes an intelligent evaluation method based on semantic feature extraction. The method design and validation were completed across four aspects: multi-source corpus organization, prompt-aware representation construction, joint scoring, and deployment interfaces.

(1) This paper constructs a multi-source sample organization framework for English writing evaluation, achieving score normalization, label alignment, and five-dimensional meta-trait mapping across ASAP++, ICLE++, DREsS, and PERSUADE 1.0. This enables comprehensive scoring, fine-grained t- -trait scoring, and text-based auxiliary supervision to operate within a unified training space.

(2) We put forward a prompt-related semantic feature obtaining and multi-angle combined marking model that combines overall semantics, sentence-layer matching, text structure arrangement, and language proof materials into one unified marking frame, thus promoting the logical consistence between the total score and trait marks through consistency restriction conditions. The outcomes make manifest that this method has better performance than representative baseline methods on three aspects, which are total scores, cross-task generalization ability, and stability in different dimensions, therefore it shows particularly strong and firm performance in low-resource and EFL environments.

(3) This paper also admits the current restrictions of the method: the description of deep argument sufficiency, extreme cut samples, and fine-grained syntax mistakes still stays not

enough. In the future, work could moreover further add argumentative structure expressions, more detailed wrong type modeling, and stricter fairness and uncertainty calibration methods to raise its application scope in the real teaching and examination situations.

About the Author

Pingping Liu was born in Quanzhou, Fujian, P.R. China, in 1988. She obtained a master's degree from Huaqiao University in China. She is currently teaching at the School of Education, Quanzhou Vocational and Technical University. Her primary research focus is English language teaching.

References

[1] Li, S., & Ng, V. (2024). Automated essay scoring: Recent successes and future directions. In Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (pp. 8114-8122).

[2] Li, S., & Ng, V. (2024). Automated essay scoring: A reflection on the state of the art. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 17876-17888).

[3] Wang, Q. (2024). A multifaceted architecture to automate essay scoring for assessing English article writing: Integrating semantic, thematic, and linguistic representations. Computers and Electrical Engineering, 118(Part A), 109308.

[4] Pack, A., Barrett, A., & Escalante, J. (2024). Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. Computers and Education: Artificial Intelligence, 6, 100234.

[5] Li, S., & Ng, V. (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) (pp. 7661-7681).

[6] Li, S., & Ng, V. (2024). ICLE++: Modeling fine-grained traits for holistic essay scoring. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 8465-8486).

[7] Yoo, H., Han, J., Ahn, S.-Y., et al. (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) (pp. 13439-13454).

[8] Do, H., Park, T., Ryu, S., et al. (2025). Towards prompt generalization: Grammar-aware cross-prompt automated essay scoring. In Findings of the Association for Computational Linguistics: NAACL 2025 (pp. 2818-2824).

[9] Chen, P.-K., Tsai, B.-W., Wei, S. K., et al. (2025). Mixture of ordered scoring experts for cross-prompt essay trait scoring. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 18071-18084).

[10] Chu, S. Y., Kim, J. W., Wong, B., et al. (2025). Rationale behind essay scores: Enhancing S-LLM's multi-trait essay scoring with rationale generated by LLMs. In Findings of the Association for Computational Linguistics: NAACL 2025 (pp. 5811-5829).

[11] Eltanbouly, S., Albatarni, S., & Elsayed, T. (2025). TRATES: Trait-specific rubric-assisted cross-prompt essay scoring. In Findings of the Association for Computational Linguistics: ACL 2025 (pp. 20528-20543).

[12] Li, S., & Ng, V. (2025). Graph-based multi-trait essay scoring. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (pp. 33325-33351).

[13] Kwako, A., & Ormerod, C. (2024). Can language models guess your identity? Analyzing demographic biases in AI essay scoring. In Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024) (pp. 78-86).

[14] Karim, A., Wang, Q., & Yuan, Z. (2025). Beyond the score: Uncertainty-calibrated LLMs for automated essay assessment. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (pp. 19631-19636).

[15] Crossley, S. A., Baffour, P., Tian, Y., et al. (2022). The persuasive essays for rating, selecting, and understanding argumentative and discourse elements (PERSUADE) corpus 1.0. Assessing Writing, 54, 100667.

[16] Mathias, S., & Bhattacharyya, P. (2018). ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores. In Proceedings of the 11th International Conference on Language Resources and Evaluation (pp. 1169-1173).

[17] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (pp. 3982-3992).

[18] Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. arXiv, arXiv:2004.05150.

[19] Farag, Y., Yannakoudakis, H., & Briscoe, T. (2018). Neural automated essay scoring and coherence modeling for adversarially crafted input. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 263-271).

[20] Uto, M., Xie, Y., & Ueno, M. (2020). Neural automated essay scoring incorporating handcrafted features. In Proceedings of the 28th International Conference on Computational Linguistics (pp. 6077-6088).

[21] Phandi, P., Chai, K. M. A., & Ng, H. T. (2015). Flexible domain adaptation for automated essay scoring using correlated linear regression. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (pp. 431-439).

[22] Wang, Y., Wang, C., Li, R., et al. (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 (pp. 3416-3425).

[23] Kumar, R., Mathias, S., Saha, S., et al. (2022). Many hands make light work: Using essay traits to automatically score essays. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 1485-1495).

[24] Do, H., Kim, Y., & Lee, G. G. (2023). Prompt- and trait relation-aware cross-prompt essay trait scoring. In Findings of the Association for Computational Linguistics: ACL 2023 (pp. 1538-1551).

[25] Chen, Y., & Li, X. (2023). PMAES: Prompt-mapping contrastive learning for cross-prompt automated essay scoring. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1489-1503).