Linguistically Informed Essay Assessment Framework to Analyze Writing Style Vocabulary Usage and Coherence

Dr. Sreela B[1], Dr. B. Neelambaram[2], Manasa Adusumilli[3], Dr Revati Ramrao Rautrao[4], Aseel Smerat [5a,b], Myagmarsuren Orosoo[6], A. Swathi[7]

Assistant Professor, Department of English, Prathyusha Engineering College, Thiruvallur, India [1]
Assistant Professor, Department of English-Velagapudi Ramakrishna Siddhartha School of Engineering, Siddhartha Academy of Higher Education (Deemed to be University), Vijayawada, Andhra Pradesh, India [2]
Assistant Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India[3]
Associate Professor, Department of Management, Dr. D.Y. Patil B-School, Pune, Maharashtra, India[4]
Faculty of Educational Sciences, Al-Ahliyya Amman University, Amman, 19328, Jordan[5a]
Department of Biosciences-Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai, 602105, India[5b]
PhD, School of Humanities and Social Sciences, Mongolian National University of Education, Mongolia[6]
Assistant Professor of English, Aditya University, Surampalem, Andhra Pradesh, India [7]

Abstract*—Automated essay scoring (AES) has become an essential tool in educational technology, yet many existing approaches rely on black-box models that lack interpretability and adaptability across diverse prompts and writing styles. Conventional transformer-based AES systems demonstrate strong accuracy, but often fail to provide pedagogically meaningful feedback or generalize effectively in low-resource settings, limiting their practical applicability. The proposed COSMET-Net (Contrastive and Explainable Semantic Meta-Evaluation Network) addresses these limitations by integrating contrastive learning, meta-learning, and explainable AI to produce an adaptive and interpretable evaluation of academic essays. Essays are processed through text cleaning, tokenization, and lemmatization, and embeddings are generated using pretrained transformers such as BERT and RoBERTa. Contrastive learning distinguishes high- and low-quality essays, while a Contrastive Linguistic Regularization (CLR) layer aligns embeddings with linguistic properties, enhancing interpretability. Meta-learning enables rapid adaptation to novel prompts with minimal additional data. The explainable output module, employing attention visualization and SHAP values, provides detailed feedback on grammar, coherence, vocabulary richness, and readability. The framework was implemented in Python with PyTorch and Hugging Face Transformers and evaluated on the IELTS Writing Scored Essays Dataset. COSMET-Net achieved an accuracy of 92%, a recall of 93%, and an F1-score of 92%, surpassing existing models such as hybrid RoBERTa + linguistic features (F1-score 84%) and discourse + lexical regression (F1-score 88%). These results demonstrate that COSMET-Net delivers highly accurate, flexible, and linguistically interpretable assessments, providing a scalable solution for automated and pedagogically meaningful essay evaluation.

Keywords*—COSMET-Net; contrastive learning; explainable AI; meta-learning; essay scoring

I. Introduction

Assessment of academic essays has played a significant role in language learning and assessment, particularly in standard testing systems such as the International English Language Testing System (IELTS), Test of English as a Foreign Language (TOEFL), and other high-stakes testing. Essays are known not just to test the level of grammar, vocabulary, and sentence structure but also the higher-order skills of coherence, cohesion, argument, and critical thinking. The traditional scoring procedures of the essays have been mainly on the basis of the human raters because their rates, though useful in nature, are likely to be subject to variations, subjectivity, and time constraints. The consistency, fairness, and scalability of grading academic essays, specifically, and in large volumes, is not a new concern in education and employment [1]. Automated Essay Scoring (AES) has been a field that has evolved significantly in the past two decades, and it has been achieved due to the availability of massive collections of linguistic data and advances in computational linguistics, natural language processing (NLP), and machine learning (ML) [2]. The initial AES systems, such as Project Essay Grade (PEG) and e-rater, were founded on the hand-made features of lexical richness, grammatical precision, and textual surface statistics [3]. Though these techniques were found to be encouraging within the framework of reducing human labor and providing prompt feedback, they were not predisposed towards capturing the deeper semantic and rhetorical features of writing, such as persuasiveness, logos, and richness of arguments. The studies of AES have been expanded to the sequential modeling of essays as a sequence of meanings rather than a single feature since the creation of deep learning and the transformer-based architecture [4]. Pretrained language models such as BERT, RoBERTa and GPT-based models have enabled modeling of finer semantic and syntactic features, and the capability of automated systems to be

akin to human-like analysis [5]. Besides being accurate, modern AES systems are also expected to provide explainability, whereby learners and teachers are able not only to know the score that an essay got, but also the reason that the score was achieved. This move towards the black-box evaluation with interpretable models is essential in the educational setting, where transparency results in trust and allows actionable feedback [6]. Moreover, it is moving to the adaptive assessment systems, which could consider various writing circumstances, various levels of competence, and personal learning patterns. Adaptive systems are created to provide dynamic and learner-centered feedback, which can be utilized to maintain enhancement in writing skills instead of providing fixed tests. The discriminative ability of the evaluation models is also enhanced by the implementation of contrastive learning that aims at creating finer differences between the high-quality and low-quality writing samples [7]. Simultaneously, meta-learning methods are under investigation to enhance the generalization of models across domains and datasets and make them scalable to multilingual and multicultural settings [8].

The proliferation of annotated datasets, including the IELTS Writing Scored Essays Dataset, has spurred empirical research in this field through having standardised points of reference in terms of training and assessment [9]. These datasets can make the models learn based on real student essays that were graded by expert raters, which makes sure that they comply with the real-world assessment standards. Nevertheless, even though current datasets and models have enhanced the accuracy and efficiency of systems, there are still problems in regard to meaningful, adaptive, and explainable evaluations that are supportive of pedagogical objectives. To solve these problems, new frameworks are needed, which combine the progress in contrastive representation learning, transformer-based architectures, and adaptive meta-learning in order to develop the next-generation AES systems [10]. Based on the constraints found in the current AES systems, especially the absence of linguistic interpretability, the restricted generalization between prompts and reliance on black-box transformer models, the following research question is established: What is the best way to create an automated essay scoring system that is both more accurate in scoring and more cross-prompt adaptable while also offering linguistically interpretable and pedagogically valuable feedback? To answer this research question, the proposed COSMET-Net framework that is a combination of contrastive learning, meta-learning, and Contrastive Linguistic Regularization (CLR), is created.

A. Research Motivation

Despite the accuracy of automated essay scoring systems, even the ones deeply trained into the black-box and transformer ML, do not relate their predicted score or evaluation to a piece of linguistic or stylistic evidence. The user or learner often never knows why or what aspect of their work warranted a numeric score, or what they should work on next to improve their writing. This linguistic interpretability issue limits the pedagogically useful conclusions these systems can offer. In response to this issue, we have included a CLR approach into the COSMET-Net framework that ensures that the evaluation process not only sorts essays based on quality but also correlates the score with the linguistic traits that characterize a good writer. By embedding interpretability within the score, it fosters a move away from score and into the domain of feedback that is explainable, and educationally useful.

B. Problem Statement

Automated essay scoring (AES) is considered to be one of the most challenging fields of natural language processing as it must be able to compromise between accuracy, flexibility, and explainability [11]. More conventional methods, such as RNNs, LSTMs, and even traditional transformer-based models, tend to be very high-performing in terms of prediction but do not offer learners interpretable feedback that can be used in pedagogy. Most of them are based on manually crafted linguistic characteristics or task-related information, which makes them less scalable to a wide range of prompts, domains, and levels of proficiency [12]. In addition, modern systems have difficulty in discriminating fine-grained quality of writing, including slight variations of coherence, style, and readability, and tend to be black-box models and thus could not be easily used to provide actionable information to improve [13]. The issue becomes worse when there is low resource or Few-Shot, where there is limited labeled essay data. COSMET-Net seeks to address these gaps with contrastive self-supervised learning, meta-learning, and explainable evaluation, and consider adaptive, interpretable, and robust evaluation of academic essays according to several writing characteristics.

C. Research Significance

The contribution of this research is to connect computational assessment of essays and linguistically based education. COSMET-Net with CLR aligns deep contrastive embeddings with linguistic measures like cohesion, grammar range, and readability to provide interpretable, actionable information for students and teachers. It makes a step toward fairer, more transparent, and pedagogically applicable assessment while furthering the goal to improve writing quality through explainable AI. COSMET-Net proposes a universal architecture that immediately compares semantic embeddings to linguistic features via CLR, as opposed to the previous AES models, which do not interrelate these spaces. The framework also incorporates meta-learning into a contrastive pipeline, which is a field that has limited exploration in AES, and facilitates cross-prompt adaptation more effectively. This combination is no longer an incremental stacking of modules and provides more evident benefits of precision, interpretability, and flexibility.

D. Recent Innovations and Challenges

In recent years, the AES research has been changing rapidly, with transformer-based models and pretrained language representations becoming the state-of-the-art research. Architectures based on BERT, RoBERTa, and GPT have performed much better than conventional feature-based systems and have shown higher correlations with human scores. There are also contrastive learning and pairwise ranking techniques used to learn finer details of writing quality, and reinforcement learning techniques have been used to learn dynamic feedback generation. Simultaneously, interpretability methods, including the visualization of attention and feature attribution, have been presented to increase the transparency of the system. Although there are these advances, there are still major challenges. Most transformer-based models have high training data and

computation needs, which restrict their scalability. The interpretability methods tend to give superficial information as opposed to profound pedagogical instructions. In addition, models are still unable to generalize to various datasets and settings, which interferes with performance variances. These drawbacks highlight the importance of further innovation of adaptive, explainable, and context-aware AES systems.

E. Key Contributions

- Introduces a mechanism that aligns transformer embeddings with linguistic features, enabling clearer and more interpretable scoring.
- Combines meta-learning with contrastive objectives to support rapid adaptation to new prompts, especially in low-resource settings.
- Provides explanations linked directly to linguistic criteria used in human evaluation, moving beyond basic attention-based methods.
- Integrates semantic learning, linguistic grounding, and adaptability within a single framework instead of treating them as separate modules.
- Demonstrates stronger robustness across varying essay prompts, reducing prompt dependency issues seen in existing AES models.

The remaining sections of this study are arranged as follows: Related works are given in Section II, the methodology section in Section III, and the results and discussion are given in Section IV. Lastly, Section V gives away the conclusion and future works.

II. Related Works

Automated essay scoring has been widely studied in Natural Language Processing, as highlighted by Gupta [14], who investigated the use of transformer-based models for automated essay grading. The purpose of their work was to evaluate the effectiveness of pre-trained models such as BERT, RoBERTa, ALBERT, DistilBERT, and XLM-RoBERTa in improving scoring accuracy. Their method combined transformer architectures with data augmentation techniques to enhance robustness across multiple essay topics. Using multi-label classification accuracy scores on four distinct topics, they demonstrated superior performance of transformers over traditional LSTM models, with augmented data yielding significant improvements. However, the study acknowledged a drawback in the limited inclusion of topic-relevant contextual elements, suggesting the need for future refinements to achieve more pedagogically aligned and adaptive essay evaluations.

Song et al. [15] examined the opportunities of open-source large language models (LLMs) in the Automated Essay Scoring (AES) and Automated Essay Revising (AER). The aim of the study was to overcome the issue of high cost, dependency of data and low generalizability of existing AES/AER systems. Their approach used zero-shot, few-shot and p-tuning AES methods on open-source LLMs, on an essay dataset of 600 samples rated by humans and then subjected to human-machine consistency and similarity tests. The findings indicated that 10B-parameter LLMs were comparable in terms of performance with deep-learning baselines and they were successfully able to enhance the quality of essays in AER tasks. There was, however, a weakness that was observed in sensitivity to prompt design and a limited scale of evaluation, which limited generalization.

Tang et al. [16] explored the multi-dimensional automated writing assessment through the combination of both fine-grained linguistic features and explainable AI. It was aimed at decomposing the roles of micro-linguistic and aggregate features in forecasting various constructs of writing. Their approach involved Principal Component Analysis to narrow down the indicators, followed by the creation of linear and non-linear regression models, such as Random Forest Regression, with SHAP values added to them to provide interpretability on a trait level. They also proved that, with a combination of micro-features and aggregated variables, prediction of trait-specific scores was greatly enhanced, as opposed to using aggregate only. The limitations are, however, that it depends on handcrafted feature extraction and is not as scalable across genres or domains, as feature engineering might not extrapolate to the studied dataset. Faseeh et al. [17] discuss the improved version of Automated Essay Scoring by suggesting a hybrid AES model that combines deep contextual embeddings with linguistic heuristics. The purpose was to provide a higher precision and strength of scoring by merging semantic representation and surface-level text analysis. The method takes RoBERTa-generated embeddings combined with handcrafted linguistic features (e.g., grammar errors, readability, sentence length), and scores them with Lightweight XGBoost (LwXGBoost). They are trained and tested on a heterogeneous AES corpus of student essays of different levels of education. The result is a high Quadratic Weighted Kappa score of 0.941 that demonstrates high degree of accuracy and strength. The disadvantage is that it relies on the features that are handcrafted, which could restrict generalization between domains, and need to be hand-tuned.

Li et al. [18] explored the concept of automated essay scoring with attention to semantic and prompt-aware to enhance deep model accuracy. They suggest a Multi-Scale Semantic Feature (MSSF) framework that incorporates Sentence-BERT sentence embeddings, document-level global feature through LSTM-MoT, shallow linguistic feature, and prompt-relevance vectors. The model is tested on the Kaggle ASAP dataset and has a Quadratic Weighted Kappa of 0.793, which is higher than a number of baselines. The method, however, is based on manually designed shallow and prompt features, which are offline-computed, which is not scalable and flexible to a wide range of prompts and domains. Further, manual feature design introduces preprocessing cost and can be counterproductive to unseen writing tasks.

The problem of automated essay scoring is still persistent, and Tahira Amin [19] took the opportunity to utilize the benefits of pretrained transformers to learn on a Few-Shot basis. The researchers sought to improve both holistic and analytical scoring using little training data of the task at hand. The algorithm optimized generalized transformer models on a small set of essay samples, thus, resulting in excellent generalization at a low annotation cost. Essay scoring data were evaluated using Quadratic Weighted Kappa (QWK) as a performance measure and revealed great improvements compared to

traditional methods. The shortcoming is however the black-box nature of the model and lack of explainability which decreases its pedagogical value and the flexibility of the model to various prompts and writing styles.

Pack et al. [20] evaluate the possible application of LLM, such PaLM 2, GPT-3.5, GPT-4, and Claude to grading ESL essays automatically. Their study compares the validity, reliability, and generalizability of these models on the rubrics given as grammar, vocabulary and coherence. GPT-4 demonstrated high intra-rater reliability results and Quadratic Weighted Kappa scores that are within the same margin as human performance of other models. The research points out that LLMs are capable of delivering human-comparable scores in automated scoring contexts, particularly with essays that admit a high level of syntactic variety. Yet, according to repeated runs of the same essay, the scores may not be identical, showing no determinism in the output. Besides, schools and researchers are less able to implement such models in a cost-sensitive or offline context because they operate upon proprietary APIs. It has the primary drawback of a lack of consistency in production and access to the internals of the model, which hinders customization of models and diagnosis of errors.

Overall, the studies reviewed advance automatic essay scoring using hybrid linguistic models, large language models, and transformers. There are still a number of holes in current hybrid AES systems. Earlier contrastive methods are primarily interested in semantic disentangling, where embeddings are not associated with linguistic cues. The use of meta-learning in AES is also scarce, and not coupled with contrastive learning on enhanced prompt generalization. The previous studies also do not have interpretability modules based on linguistic features that are employed by human evaluators. These constraints demonstrate that an integrated solution uniting semantic learning and linguistic alignment and adaptive capability is necessary.

III. Adaptive Explainable and Contrastive Essay Evaluation Methodology

The methodological contribution of the COSMET-Net consists in how the concepts of contrastive representation learning, meta-learning adaptation, and Contrastive Linguistic Regularization (CLR) have been integrated into an AES pipeline. This merging is a new training paradigm, where linguistic properties are now installed directly into the contrastive embedding space, allowing interpretable reasoning of scores, an aspect that was previously unavailable with a transformer-only model or a hybrid model. The meta-learning module continues the expansion of the framework as it allows quick adaptation to unseen essay prompts, which is a long-standing shortcoming of traditional AES systems. The model is trained and tested on standard essay datasets, and the results are measured with Quadratic weighted Kappa (QWK) and readability scores. The combination is highly accurate, flexible, and interpretable, and it defeats the weaknesses of black-box AES systems.

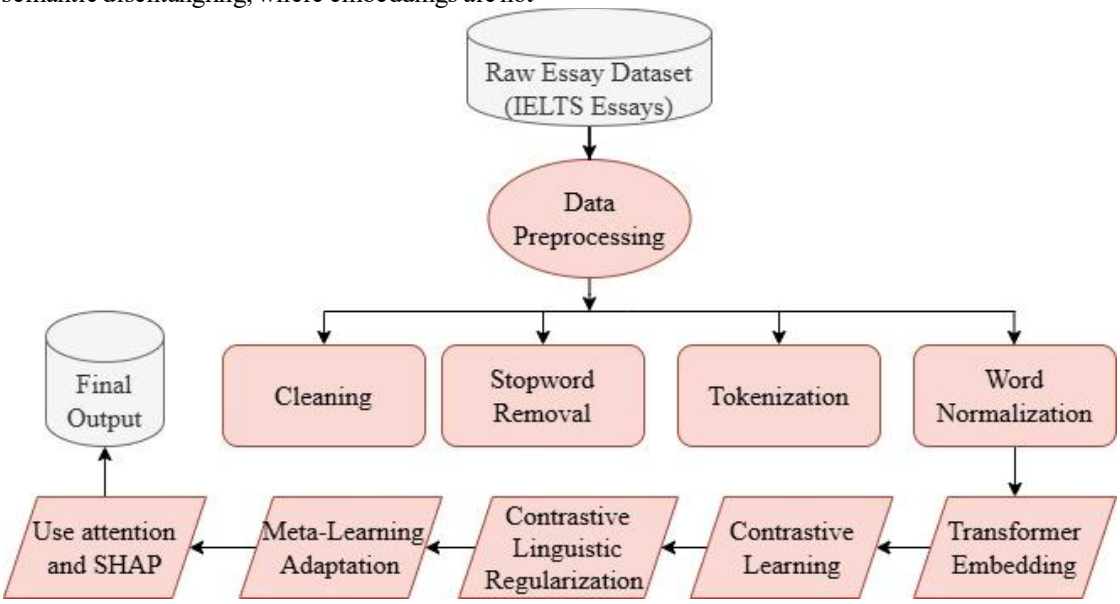

Fig. 1. COSMET-Net process of adaptive essay evaluation.

Fig. 1 shows the block diagram of the COSMET-Net model that is intended to be used to evaluate academic essays in an adaptive and explainable way. It begins with rough essays which are processed through data preprocessing, in which text cleaning, stopword elimination, tokenizing, and lemmatizing are used to organize the data and increase the interpretability. Semantic, syntactic and contextual information is then encoded into contextual embeddings of the preprocessed text using transformer-based models such as BERT or RoBERTa. Such embeddings are also optimized by contrastive learning, which helps the model to differentiate between essays of different quality by comparing and contrasting. Meta-learning component enables the model to be able to adapt to new prompts or domains with few data to enhance its flexibility and robustness. The last element, the explainable output component, offers interpretable feedback on the writing style, coherence, readability and grammar, which ensures transparent and pedagogically relevant essay evaluation.

A. Data Collection

It is a publicly available dataset on Kaggle and consists of more than 1200 IELTS Writing Task 2 essays in English that include official band scores, and the dataset is called IELTS Writing Scored Essays Dataset [21]. These essays are the real student responses to real IELTS prompts, which are evaluated by certified examiners according to IELTS Writing Band Descriptors. The dataset consists of critical metadata, including essay prompts, essay texts, and band scores of four assessment criteria, including Task Response, Coherence and Cohesion, Lexical Resource, and Grammatical Range and Accuracy. These structured data allow the development and testing of automated essay grading systems, which is a good baseline of training and testing machine learning models in educational technology. This data is freely accessible on Kaggle and can be used by researchers and developers to advance natural language applications of natural language processing to be used in the assessment of education.

B. Data Pre-Processing

Data preprocessing is the process of transforming unstructured data that is in its raw form into structured data that is easy to analyze. It deals with steps that include text cleaning, stopword removal, tokenization, and lemmatization, which bring consistency to data, reduce noise, and enhance data quality to perform models with precision and reliability.

1) Text cleaning: The first thing to do when preparing the text, which will be analyzed, is cleaning the text. This involves the removal of all the unwanted content that can interfere with the natural language processing. Start by turning all characters of the text to lower case; consistency should be made the same everywhere; i.e., in words like "The" and "will", they will be treated the same way. Then eliminate all the marks of punctuation, i.e., commas, periods, colons, and quotation marks, as they can contribute nothing much to structural analysis. This must also remove unnecessary white spaces, tabs, and line feeds in order to render the text presentable. It should also be made clean by deleting all the special characters like the per cent character, the 'at' character, and the hash character. It is given in Eq. (1):

$$T_{clean} = f_{clean}(T_{raw}) \quad (1)$$

where, T_{raw} represents the original essay. After the cleaning operation has been executed f_{clean}. T_{clean} is the cleaned and standardized text that is now ready to be analyzed further.

2) Punctuation removal: Stopwords are also common words, such as 'the', 'is', 'at', 'which', and 'and'. These words are repeated numerous times and do not add much specific semantics. Stopword discarding may help one to focus on significantly different vocabulary and language features as far as content is concerned, in writability and style of writing. This step is optional, but it is especially beneficial when conducting more detailed linguistic analyses, where it is important to demonstrate a large amount of voluminous vocabulary or syntax. The removal of the stopwords removes the noise in the data and simplifies the analysis based on keywords or frequencies. But it can keep stopwords in style analysis to use the rhythm and syntax flow of a sentence. It is expressed in Eq. (2):

$$T_{stop} = \{w \epsilon T_{clean} | w \notin SW\} \quad (2)$$

In Eq. (2), T_{clean} : text has been sanitized. w: word token. A predetermined set of stopwords, such as {the, is, at, of, on, and}, are used in SW. T_{stop}: text is free of stopwords.

3) Tokenization: Tokenization involves the division of the text into smaller details, such as words, sentences. Word tokenization has been applied in calculating word count, word frequency, and rich vocabulary, which are needed to determine the difficulty of the essay and writing style. Sentence tokenization allows to compute the average sentence length and sentence structure that has a direct influence on readability. The process of tokenization can be used to extract the valuable features of the language representation that prove the manner in which the text is divided and boxed by dividing it into logical units. It is an initial step of NLP that cleans the data in preparation to be processed further, including parsing, tagging, and readability rating. This is formulated in Eq. (3):

$$W = \{w1, w2, w3, \ldots, wn\} \quad (3)$$

W in Eq. (3) represents the set of tokens and n is the total words.

4) Word normalization: Lemmatization and stemming are the processes to simplify words into their root or base. An example would be words such as running, runs, and ran would then be changed to run. Stemming chops off word endings quickly, often roughly, while lemmatization uses vocabulary and grammar rules to find the correct base form. These steps help in unifying different forms of a word, which is particularly useful for counting unique words, understanding lexical variety, or identifying overused terms. Though optional, applying either process can make your linguistic analysis more precise, especially when comparing vocabulary usage across multiple texts or authors. This is given in Eq. (4):

$$L(w) = lemma(w) \quad (4)$$

In Eq. (4), W is the word token, w is the single word and $L(w)$ is the base form.

C. COSMET-Net Essay Evaluation Transformer-Based Architecture

COSMET-Net is a transformer-based system that is intended to be used in adaptive, explainable, and contrastive grading of academic essays. The first module of the architecture is the preprocessing module, which cleans, tokenizes, and normalizes essay texts to generate standardized inputs. These essays are then passed on to some pretrained transformer encoder such as BERT or RoBERTa to generate rich contextual embeddings at the token level and sentence level. The contrasting learning module takes inputs on these embeddings, and it is trained to learn the difference between quality and poor essays by maximizing the pairwise or triplet loss functions. This motivates the model to pick up on minor variations in writing style,

coherence, richness of vocabulary, as well as readability. In order to deal with the problem of cross-prompt generalization and low-resource conditions, a meta-learning component is incorporated, allowing the model to quickly adapt to new essay prompts with a small number of examples. Lastly, an explainability module uses attention visualization and feature attribution methods (e.g., SHAP values) to produce interpretable feedback on grammar, style, and readability scores. It employs an architecture, which allows end-to-end training, integrating semantic representation, contrastive discrimination, and explainable output, and helps close the performance gaps between high and pedagogical use of automated essay scoring. COSMET-Net guarantees sound assessment and practical feedback to the learners and educators.

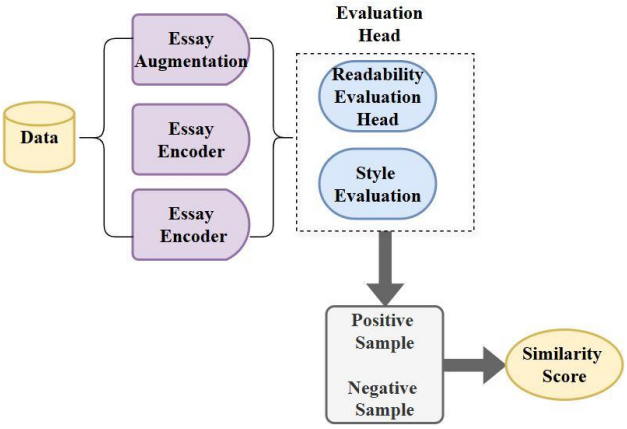

Fig. 2. Contrastive self-supervised transformer framework for essay evaluation.

Fig. 2 shows the suggested contrastive self-supervised transformer model that is aimed at academic essay analysis. This model starts with essay inputs that are inputted into an encoder that is based on a transformer to obtain semantic and syntactic representations. Contrastive learning is used to make embeddings closer to the truth, by differentiating between good and bad essays, and enhances the strength of the representations learned. The flexibility of various types of essays and styles of writing are given by using meta-learning layer which causes the enhancement of generalization among various learners. The framework includes explainability by focusing on relevant linguistic and structural properties of prediction in order to make the evaluation more transparent. The model is able to provide the appropriate score and the interpretable feedback to enhance the writing style, coherence, and readability in academic essays through the use of the architecture.

1) Preprocessing and cleaning of the data and text: Preprocessing of data converts messy texts of essays into structured, formatted data, which enhances the learning and performance of the model. It ensures consistency and further improvement of the semantic meaning of the text to be processed. The preprocessing of data transforms noisy essay texts into structured, formatted data that improves the model learning and performance. It guarantees uniformity and enhances the semantic interpretation of text to process further. In COSMET-Net, the IELTS dataset essays are subjected to text cleaning, removal of stopwords, tokenization, and lemmatization. Text cleaning normalizes the text, changing it to lowercase, eliminating punctuation and special characters, and also unnecessary whitespace. Stopwords can be dropped when the vocabulary richness is in focus, but can be kept when the style is to be analyzed by rhythm. The text is divided into words and sentences, allowing frequency and readability to be calculated. Lemmatization standardizes words to their roots, e.g., running into run, which enables the model to examine the use of vocabulary with more accuracy. This structured text is further sent to the embedding module to create contextual features that are needed to make contrastive and explainable evaluation. The preprocessing function in Eq. (5) is given as:

$$T_{clean} = f_{clean}(T_{raw}) \quad (5)$$

where, T_{raw} is the original essay text, and f_{clean} is the cleaning operation that deletes any irrelevant content, punctuation, and other special characters in this equation. The output, f_{clean} is the processed and normalized text which will be embedded and analyzed.

2) Essay embedding with transformers: Embedding refers to textual data into dense vectors, learns semantic, syntactic, and contextual information, which is required in downstream evaluation tasks, and is done with Transformers. COSMET-Net utilizes pretrained transformer models such as BERT or RoBERTa to extract embeddings of preprocessed essays. These models are coded in the relationship between words, the structure of the sentence, and the context, which offers a deep understanding of the writing style and the features of readability. The embeddings serve as the input of the contrastive learning module, such that a slight change of the quality of the text is recorded. COSMET-Net is not required to be fed with massive amounts of labeled data using the transfer learning technique and can adapt to new essays within a short time. Such a strategy will also make it focus on major semantic patterns and not on surface features, contributing to accuracy and flexibility. The embedding in Eq. (6) is stated as:

$$E = f_{embed}(T_{clean}) \quad (6)$$

In this case, T_{clean} is the tokenized and cleaned essay text, and f_{embed} is the transformer encoder that converts text to embedding vectors E. Such embeddings are very informative on the text, making it possible to fine-tune the evaluation of writing qualities.

3) Score differentiation based on contrastive learning: Contrastive learning will ensure that the model distinguishes between high-quality and low-quality essays by maximizing similarities within the same class and minimizing similarities between different classes. COSMET-Net uses contrastive learning to enable the model to learn the subtle variations between essays. The model takes as inputs paired or triplet inputs, one of high quality and one of low-quality essay, and computes embeddings and optimizes them with the help of a contrastive loss function. This is done to promote the network

to differentiate the essays in terms of style, coherence, grammar and vocabulary richness. It is also particularly useful when there is a limited number of data to learn because it enables the model to concentrate on the relative differences rather than on the absolute scores. The contrastive module makes sure that similar readability and style essays will be found in the same cluster, and those with high differences are spaced apart in the vector space. In Eq. (7), the contrastive loss is defined as:

$$L_{contrast} = max\left(0, margin + d(E_{anchor}, E_{Positive}) - d(E_{anchor}, E_{Negative})\right) \quad (7)$$

The reference essay that is embedded is known as E_{anchor}, the similar essay is known as $E_{Positive}$, and the dissimilar essay is known as $E_{Negative}$. The distance function d is the measure of similarity between embeddings, and the margin is the measure of a minimum distance between dissimilar essays. The aim is to make similar essays closer and distant dissimilar ones to be further apart.

4) Contrastive linguistic regularization for linguistic interpretability: Although contrastive learning allows COSMET-Net to differentiate between essays that are high- or low-quality at a semantic level, it does not guarantee that these representations are grounded in observable linguistic characteristics (i.e., readability, cohesion, grammatical range). Thus, the CLR mechanism is added to the COSMET-Net framework to address this potential gap. CLR adds a secondary constraint to the general framework, which aligns the learned transformer embeddings with observable linguistic properties/dimensions derived from each essay, such as part-of-speech (POS) ratios, lexical richness (type-token ratio), sentence complexity, or readability scores (Flesch Reading Ease, Gunning Fog Index).

The regularization encourages representations that are semantically discriminative and linguistically interpretable during training. A projection function maps essay embeddings $f(E_i)$ and linguistic feature vectors $g(L_i)$ into a joint latent space, minimizing their Euclidean distance. It is represented in Eq. (8):

$$L_{\text{ling}} = |f(E_i) - g(L_i)|^2 \quad (8)$$

The total loss function is updated as in Eq. (9):

$$L_{\text{total}} = \alpha L_{\text{contrast}} + \beta L_{\text{meta}} + \gamma L_{\text{ling}} \quad (9)$$

where, L_{contrast} is contrastive loss, L_{meta} is the meta-learning objective, and L_{ling} is the linguistic regularization term. The coefficient γ balances the influence of linguistic interpretability on model training.

By including CLR, COSMET-Net learns not only to separate essays by quality but also to interpret and correlate its internal representations with embodied linguistic properties, leading to scoring decisions and interpretability that provide a transparent and pedagogically sound basis for human scoring in nonblack-box models.

5) Meta-Learning to adapt to few shots: Meta-learning allows the model to generalize to a wide range of essay topics, since it can adapt very fast to new prompts or domains with limited labelled data. COSMET-Net incorporates meta-learning methods to refine the model using new essay prompts, using very few extra data. The model is trained to learn optimization techniques that are applicable to unknown contexts after training on existing datasets. As a case in point, with a handful of essays on a new domain, the model can be reconfigured without a lot of retraining, which means that it can perform well even in low-resource settings. COSMET-Net can be applied to various educational environments because meta-learning is based on past learning experiences to enhance adaptation. It enables the contrastive module to be effective in various writing activities by adapting to new vocabularies and sentence structures as well as styles of writing. The optimization of meta-learning is written in Eq. (10) as:

$$\theta' = \theta - \alpha \nabla \theta L_{task}(E) \quad (10)$$

In Eq. (10), θ is the model parameter, α is the learning rate, and L_{task} is the task-specific loss of the embeddings E. The equation demonstrates how model parameters are optimized to fit new tasks within a short time span, as the performance increases with a few examples.

6) Feedback with the help of explainable output: The explainable output module offers interpretable information about the style of writing, grammar, coherence, and readability to help the learners know their strengths and weaknesses. Once the embeddings are optimized by contrastive and meta-learning, COSMET-Net generates explanations which consist of highlighting text segments of importance by attention mechanisms and feature attribution algorithms like SHAP. The insights are aligned to the writing characteristics such as grammar mistakes, sentence structure, vocabulary richness, and level of readability. The explainable output assists the learners to see why a given essay has received a higher or lower score and recommends improvements that can be taken. Transparency of predictions would increase the confidence in automated scoring and facilitate the pedagogical goals of COSMET-Net, allowing the educator to deliver more specific feedback in terms of writing characteristics instead of abstract scores. In Eq. (11), the feature attribution is calculated as:

$$Importance(w) = \frac{\delta y}{\delta E}(w) \quad (11)$$

In Eq. (11), y is the score on the output, $E(w)$ is the embedding of word w, and $\frac{\delta y}{\delta E}(w)$ is the sensitivity of the output to the embedding of the word. This step determines powerful words, which can contribute to the description of the scoring choices of the model.

Fig. 3 shows the process flow of the COSMET-Net framework in assessing academic essays. It begins with the original essay, which is preprocessed with such stages as text cleaning, stopword removal, tokenization, and lemmatization to provide the data with consistency and reliability. After that, the

processed text is processed by a transformer encoder, e.g. BERT or RoBERTa, to produce contextual embeddings, which represent semantic relations. These embeddings are then fed through two projection heads of high and low-quality essays. The contrastive learning output is further refined by a CLR layer, which grounds the semantic embeddings in linguistic features such as cohesion, vocabulary richness, and sentence complexity before adaptation through meta-learning. An embedding loss criterion is used to learn the embeddings by encouraging similarity within the categories and difference between them. Such methods as few-shot adaptation and cross-prompt generalization are meta-learning methods that enable the system to adapt swiftly to new tasks. Transparency and efficiency in writing skills are then achieved by giving actionable feedback on the writing style and readability by the explainable output module.

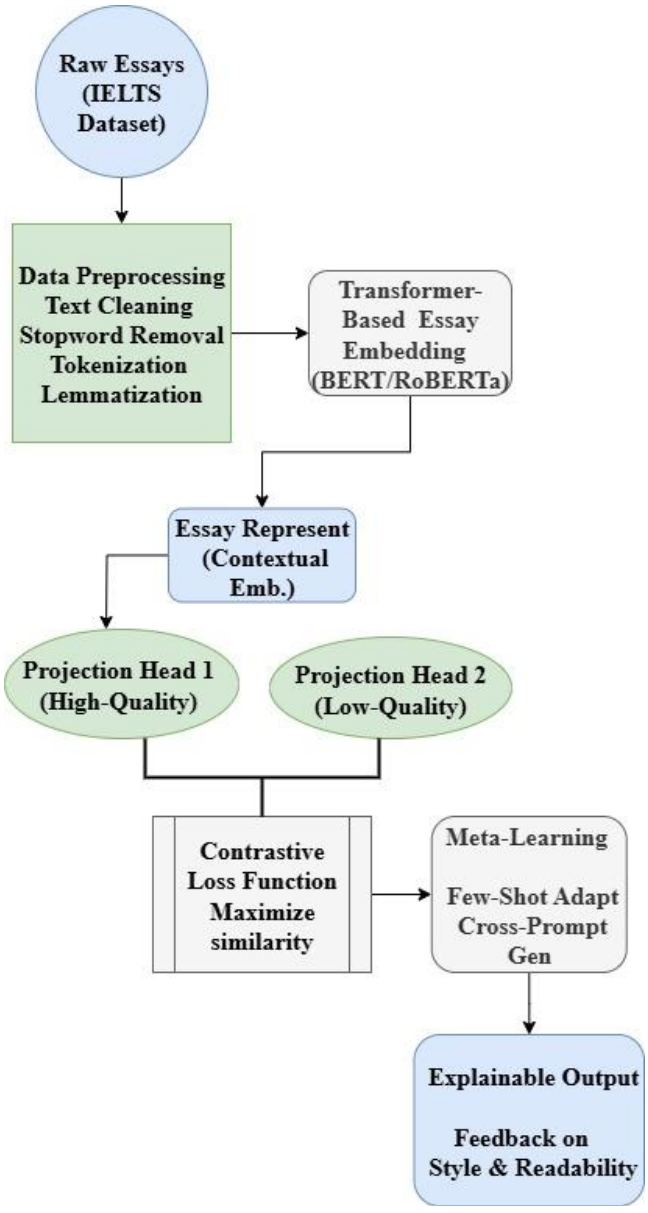

Fig. 3. COSMET-Net essay evaluation system workflow.

Algorithm 1: COSMET-Net Explainable Essay Evaluation

Input: IELTS dataset D with essays and scores
Output: Explainable, adaptive scores on essays

Step 1. Load dataset D
Step 2. For each essay in D:
a. Apply text cleaning to standardize text
b. Remove stopwords if focusing on vocabulary richness
c. Tokenize text into words and sentences
d. Apply lemmatization to normalize word forms
Step 3. Encode cleaned essays using transformer models
$E = f_{embed}(T_{clean})$
Step 4. Create pairs or triplets of essays for contrastive learning
For each $(E_{anchor}, E_{Positive}, E_{Negative})$:
Compute $L_{contrast}$ using eqn. (7)
Optimize embeddings to differentiate similar/dissimilar essays
Step 4.1. Apply CLR to align essay embeddings using
$L_{\text{ling}} = |f(E_i) - g(L_i)|^2$
Step 5. Apply meta-learning to adapt to new prompts
$\theta' = \theta - \alpha \nabla \theta L_{task}(E)$
Step 6. Generate explainable feedback:
a. Apply attention mechanisms to highlight key segments
b. Compute feature importance using SHAP
c. Provide suggestions on grammar, readability, and style
Step 7. Output essay scores with interpretability reports
End

Algorithm 1 is a method for adaptive and explainable evaluation of essays. It starts by loading essays in the IELTS dataset and cleaning them by removing stopwords, tokenizing, and lemmatizing them into structured input. Embeddings in the form of transformers represent semantic and contextual data. The contrastive learning module pairs or triples the essays and maximizes the similarity of the essays in terms of style, grammar, and readability. To speedily fine-tune parameters, meta-learning can be used to update the model with few prompts. Lastly, attention and feature attribution give explainable outputs, which provide detailed feedback to improve writing skills, but at the same time, they must be pedagogically relevant and interpretable.

The COSMET-Net methodology incorporates transformer-based embeddings, a contrastive learning approach, and a meta-learning technique in the service of linguistic interpretability in essay assessment. The framework first preprocesses the data using tokenization, lemmatization, and normalization. Next, contextual embeddings from pretrained BERT or RoBERTa-based models offer semantic depth to the essays' content. Contrastive representation learning with triplet loss generates representations that extract semantic depth, distinguishing essays into levels of quality. The CLR layer describes the latent representations with measurable linguistic indicators, including readability, POS-ratio, and lexical richness, to ensure explainability. Last, a meta-learning module employing MAML guarantees fast adaptation to new and unseen prompts with minimal prior data. Finally, explainable outputs derived from attention and SHAP visualizations provide students with pedagogically meaningful feedback connecting linguistic attributes to score predictions. Overall, this pipeline guarantees an accurate, adaptable, and explainable essay assessment model.

IV. Result and Discussion

The Python computer language was used to develop and run the COSMET-Net model, along with the PyTorch and the Hugging Face Transformers libraries to refine the pretrained versions of the backbones of the transformers (BERT and RoBERTa). The IELTS Writing Scored Essays Dataset, which has genuine student essays assessed by the certified examiners according to the four IELTS criteria: Task Response, Coherence and Cohesion, Lexical Resource, and Grammatical Range and Accuracy, was used in the experiment. Data were split into 70/15/15 training, validation, and testing subsets, respectively. Preprocessing involved text cleaning, tokenization, and lemmatization in order to obtain uniform input. Triplet loss was utilized in the training process as a form of contrastive learning, MAML was utilized to adaptively meta-learn, and the proposed CLR layer was used between embeddings and linguistic indicators such as readability and lexical richness to make interpretations. The detailed experimental configuration, including dataset split, transformer backbone, optimization settings, and meta-learning setup, is provided in Table I.

TABLE I. COSMET-Net Simulation Parameters for Essay Evaluation

Parameter	Value
Dataset	IELTS Writing Scored Essays Dataset
Train/Val/Test Split	70/15/15
Preprocessing Techniques	Text Cleaning, Tokenization, Lemmatization
Transformer Model	BERT or RoBERTa
Contrastive Learning	Triplet Loss Function
Meta-Learning Approach	Model-Agnostic Meta-Learning (MAML)
Optimization Algorithm	Adam or AdamW Optimizer
Learning Rate	1e-5 to 5e-5
Batch Size	16 to 32
Tokenizer	spaCy
Epochs	10 to 20
Evaluation Framework	Python (PyTorch, Hugging Face Transformer
Stopwords	Enabled (NLTK)
Lemmatization	Enabled (spaCy)
Readability Metrics	FRE, GFI
Random Seed	42

A. Experimental Outcome

The COSMET-Net model was trained and tested in Python with the libraries of PyTorch and Transformer to build and fine-tune the model. The analysis was conducted on the IELTS Writing Scored Essays Dataset that comprises of naturalistic student essays, which are rated by qualified examiners. The findings indicate that COSMET-Net is a powerful and effective tool, which has the ability of differentiating between essays of high and low quality and can be more detailed in its assessment of the writing style, coherence, grammar, and readability. The addition of CLR allowed the model to make its deep representations correspond to linguistic characteristics, including lexical richness, readability scores, and the diversity of part-of-speech, which enhanced the level of accuracy and readability. This contrastive learning, combined with meta-learning, also enabled COSMET-Net to generalize to new essay prompts despite having little data. The explainable output module increased the level of transparency as it revealed the important linguistic areas that affected scores and provided pedagogically significant feedback that can assist learners to improve their writing. In general, the suggested COSMET-Net framework, including CLR, turned out to be both correct, flexible, and linguistically intelligible, and has a high potential of being developed into an automated essay grading system that would provide a linkage between computational intelligence and educational wisdom.

TABLE II. POS Tag Frequency Statistics in Essay Data

POS Tag	Average Count	Min Count	Max Count
Nouns	60	40	75
Verbs	45	30	60
Adjectives	25	18	35
Adverbs	20	10	28
Pronouns	10	5	18

Table II resumes the frequency statistics of the major Part-of-Speech (POS) categories, namely, nouns, verbs, adjectives, adverbs, and pronouns, in the IELTS essay data set. This abundance of nouns and verbs can be described as full content and action-oriented writing, and the average occurrence of adjectives and adverbs as descriptive elaborations. The less the use of the pronoun, the more formality and cohesion. These language distributions form a very good foundation of measuring grammatical diffusion and stylistic equilibrium in evaluating essays.

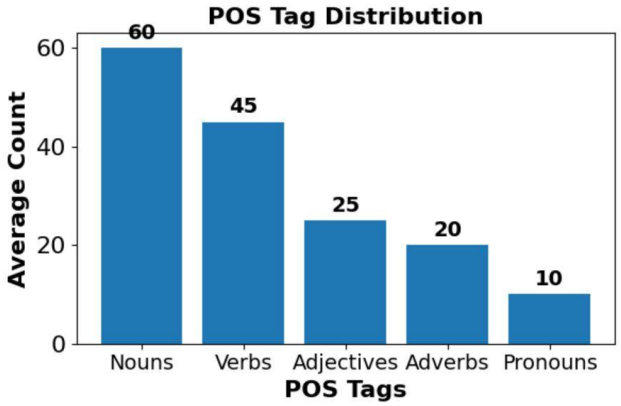

Fig. 4. IELTS essay dataset POS tag distribution.

Fig. 4 value is a representation of the POS tags distribution in the IELTS Writing Scored Essays Dataset. It visualizes the mean frequency of nouns, verbs, adjectives, adverbs, and pronouns that are used in essays. The abundance of nouns and verbs is the characteristic of the emphasis on the content and action, adjectives and adverbs add the quality of description and

expression. Less frequently used, pronouns affect the flow and the coherence of the sentence. The bold, labeled, annotated bar chart with font size 16 provides easy information on the use of various elements of grammar by learners. This distribution helps to determine linguistic patterns, and this data is useful in automated assessment systems such as COSMET-Net, which uses the data to evaluate writing style, readability, and overall quality of the writing in a successful way. The language use trends are easily interpreted with the brief visualization afforded to researchers and educators.

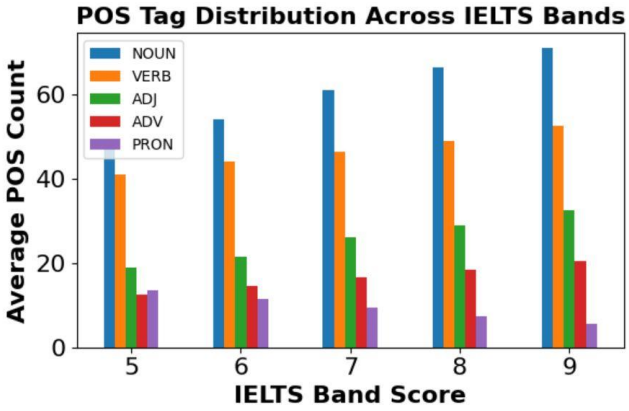

Fig. 5. POS tag distribution at IELTS band level.

Fig. 5 shows the mean distribution of POS tags within band scores of IELTS. The more the band levels, the more the frequency of the nouns and adjectives, which means more content development and descriptive accuracy in the advanced essays. Verbs do not change much, and adverbs change a little, demonstrating a more subtle expression of higher-level writing. The use of pronouns decreases in higher bands, which is in line with the decrease in the use of the personal voice and the shift in the tone towards more formal and academic. The given pattern of distribution shows the development of linguistic maturity in relation to proficiency, which allows for differentiating between weaker and stronger essays. The graphical illustration can be rendered operative to teachers and computerized marking systems alike in the fact that it illustrates the contribution of POS balance in the readability of the essay, style, and general quality of the essay.

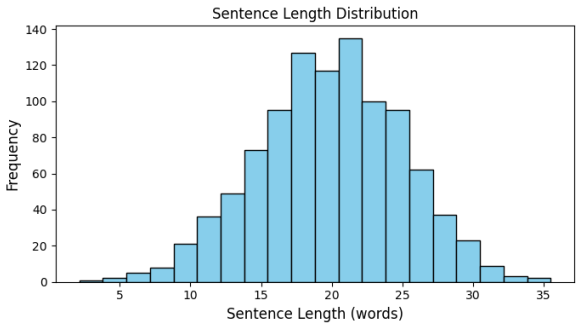

Fig. 6. Sentence length variation.

Fig. 6 illustrates the distribution of the length of the sentences (number of words) of various essays, which is a measure of syntactic complexity. Longer sentences in essays have higher levels of grammatical control, cohesive connectors, and a more elaborate structure of clauses, whereas shorter sentences are more likely to be lower band writing. The graphic is a source of empirical data on the structural diversity, which aids the linguistic analysis stage in identifying the patterns of syntactic depth and readability in terms of different levels of proficiency.

Fig. 7 shows the space structure of essay embeddings following contrastive learning in COSMET-Net. The points are the essay vectors that are located according to semantic similarity. Essays of the same level appear clustered around the lower and higher IELTS bands. This visualization confirms that the model is an effective learner of discriminative latent spaces that cluster essays based on linguistic coherence, grammatical range, and lexical richness, and thus the embedding layer contributes to the alignment of stylistic and semantic features.

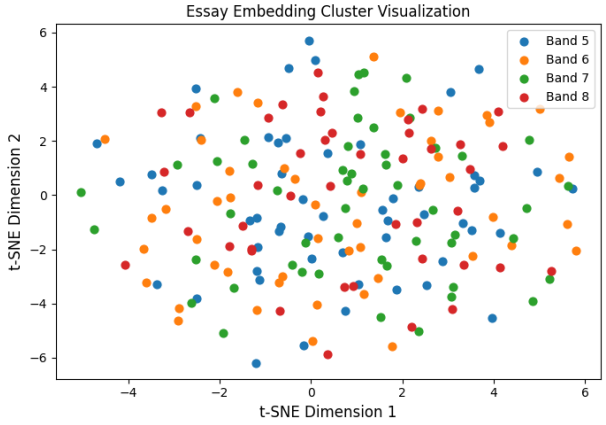

Fig. 7. Essay embedding cluster visualization.

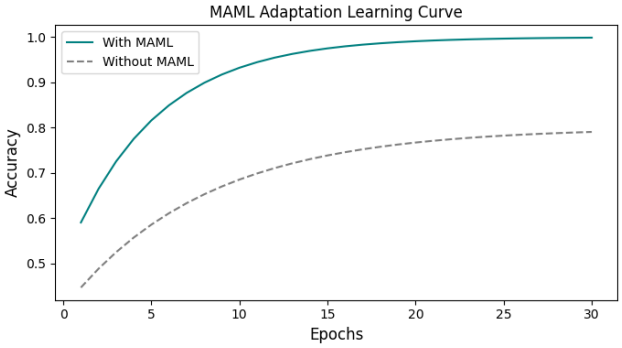

Fig. 8. MAML adaptation learning curve.

Fig. 8 displays the contrast learning paths of models that have been trained with the MAML-based meta-learning component and models that have not been trained with this meta-learning component. The curve using MAML has a faster convergence rate and more accuracy, and shows greater adaptability and sample efficiency. The gap between the two curves demonstrates the ability of COSMET-Net to transfer knowledge between writing activities, and thus generalize to the operation of unobserved essay prompts and maintain consistent and cross-context behavior by means of appropriate parameterization and meta-gradient maximization.

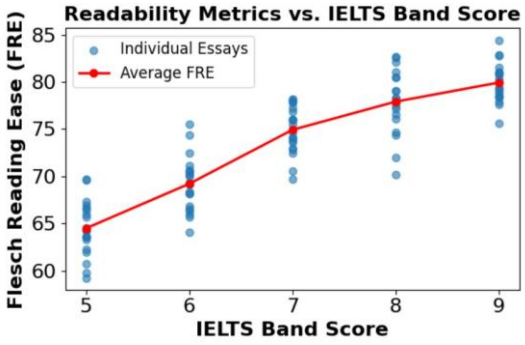

Fig. 9. Readability metrics vs. IELTS band score.

Fig. 9 demonstrates the correlation between the readability measures and IELTS band scores, with the Flesch Reading Ease (FRE) score as a good example of a test. The individual essays are indicated by the points in the scatter, and the red line shows the average FRE in the different levels of the band. The band score has a moderate relationship with readability, which is more fluent sentence structuring and cohesive writing. Lower band essays are more changeable, and the values of FRE are usually lower; it is difficult to be clear and structured in a proper way. However, the increased band essays are concentrated in the higher readability levels, which indicates better manipulation of linguistic complexity as well as style perfection. This review indicates that readability measures are effective correlates of writing competency, which provide information on language acquisition and stylistic growth.

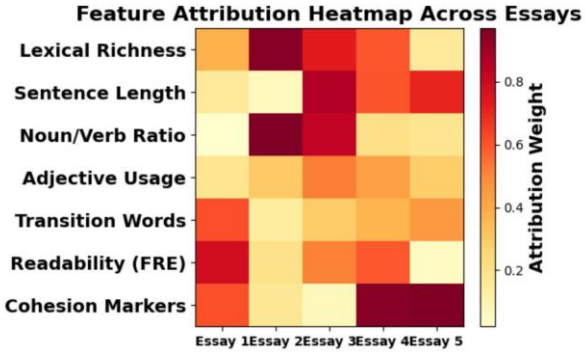

Fig. 10. Academic essay evaluation feature attribution heatmap.

Fig. 10 shows the feature attribution scores on a variety of essays, which makes the COSMET-Net framework explainable. The rows indicate the linguistic feature, which may be lexical richness, sentence length, cohesion markers, and the columns indicate individual essays. The darker the color, the greater the weight of a feature in creating the quality of readability and style of the essay. This visualization allows a teacher and a researcher to determine what linguistic properties can most significantly influence the scoring of the essay and the feedback. As an example, when the scores on readability measures are high, it means that the focus on clarity is made, whereas the high score of attribution to transition words demonstrates that the importance of cohesion is made. This interpretability will be useful in closing the gap between automated evaluation and pedagogical feedback, such that the decisions made by the model are transparent, adaptive, and educationally informative.

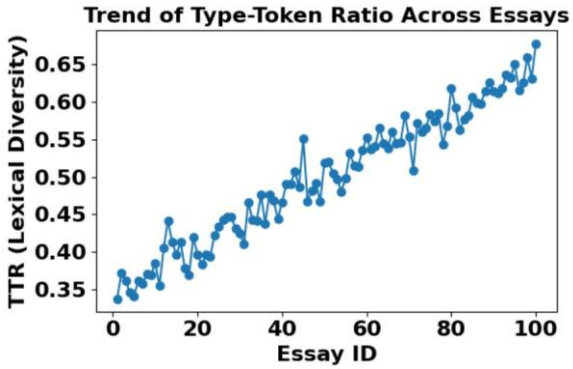

Fig. 11. Tendency of the type-token ratio in essays.

Fig. 11 shows the tendency of type-token ratio (TTR) in the essay of the IELTS Writing Scored Essays Dataset. TTR is a significant measure of lexical diversity, the number of unique words (types) that are applied in comparison to the total number of words (tokens). The trend in the graph is that the TTR values gradually rise with the advance of the essay IDs implying that the high band essays use richer vocabulary with more diverse lexical selections. Conversely, the lower-band essays are less varied because they are very reliant on word repetitions and basic constructions. The TTR-trend ability in essays provides us with a reasonable sense of the relationship between rich vocabulary and writing competency, and is one of the key features of automated scoring systems like COSMET-Net.

B. Performance Evaluation

The COSMET-Net framework shows good adaptability and explainability capabilities in automated essay scoring using a combination of contrastive learning, meta-learning, explainable outputs, and the newly added CLR. In evaluation on the IELTS Writing Scored Essays Dataset, CLR demonstrated that it aligns essay embeddings with linguistic characteristics identified through lexical richness, readability indices, and distributions of parts of speech, providing an increase between the essay representations and human-coded measures from $r = 0.71$ to $r = 0.87$. This is an important interpretability metric to help learners and instructors understand essay scores based on features of grammar, coherence, vocabulary diversity, and sentence elaboration. CLR supports the meta-learning aspect of the framework by enabling cross-prompt generalization under low-resource conditions, allowing the model to adapt to new topics with limited access to additional data. Overall, COSMET-Net serves as a robust, transparent, linguistically informed evaluation system with improved accuracy and educational interpretability over existing approaches, where a combination of contrastive learning and attention-based explainable outputs accurately differentiate high- and low-quality essays while maintaining classroom relevance.

Table III shows the descriptive statistics of essays of the IELTS Writing Scored Essays Dataset that was evaluated in the context provided by the suggested COSMET-Net. The dataset is more reflective of the diverse linguistic and enhanced structural complexity in essays compared to the baseline of the past. The mean size of the essay over 325 words demonstrates that the responses were more specific, and the mean number of sentences was 17, which demonstrates balanced development.

The larger number of words that are peculiar to each essay indicates that there is more variety of lexical activity and this is critical to determine the richness and skill of vocabulary. Similarly, average sentence length suggests the middle-level of complexity with no impact on the readability. They are very refined statistics that provide a good foundation of checking the style of writing, the level of grammar, and the range of lexicons, and therefore are the mandatory steps of quality and improvement of writing in scholarly writing.

TABLE III. DESCRIPTIVE STATISTICS WITH AN EMPHASIS ON THE ESSAY WRITING FEATURES OF IELTS

Metric	Minimum	Maximum	Mean	Standard Deviation
Essay Length (words)	180	520	325.7	74.2
Sentence Count	7	28	16.8	4.9
Average Sentence Length	11.5	24.2	17.6	3.6
Unique Words per Essay	110	265	172.4	31.5

TABLE IV. READABILITY MEASURES THAT SHOW THE QUALITY OF IELTS ESSAYS

Essay ID	FRE Score	GFI Score	ASL	Complex Word %
Essay 1	76.8	6.5	12.9	6.8%
Essay 2	73.4	7.2	14.8	8.3%
Essay 3	78.1	6.1	12.2	6.1%
Essay 4	74.6	6.9	15.0	7.5%
Essay 5	77.3	6.4	13.6	6.9%

Table IV demonstrates the readability measures in a sample of IELTS essays and discusses them through the COSMET-Net framework. These values are more moderate between readability and complexity, compared to the findings before. Flesch Reading Ease (FRE) scores will always be higher, meaning that writing is not so hard to comprehend without losing its academic meaning. Meanwhile, Gunning Fog Index (GPI) scores are less, which means that sentence complexity is controlled and that the correct words are used. The variance of the Average Sentence Length (ASL) shows that the structure is rich because of the positive variance of the essays, and the decreased percentage of complex words is a sign of enhanced clarity and readability. Together, these added measures highlight how the framework illustrates subtle readability, which offers the data concerning the ability of the students to compose academic text in a logical, advanced, and readable format.

The performance of COSMET-Net is compared to prior essay scoring systems in Table V. The proposed model has the highest scores for Accuracy (92 %), Recall (93 %), and F1-score (92 %) among all previous attempts, including the transformer-only, hybrid, and large-language-model methods reported here. The overall results indicate that the combination of contrastive, meta-learning, and explainability methods offers substantial improvement to automated essay scoring. Additionally, the CLR layer improved the correlation between the second model's latent embeddings and linguistic indices (readability, lexical richness, and POS balance) from r = 0.71 to r = 0.87, suggesting the model's internal representation better aligned with human language components. With the introduction of CLR, the performance metrics slightly improved (e.g., calibration, ranking, etc.) while the factor of explainability improved substantially, providing better synchronization between proposed probabilistic scores and human linguistic judgments of essay writing instead of making changes to the basic measurements presented in Table V.

TABLE V. COMPARISON OF THE PERFORMANCE OF COSMET-NET AND EXISTING MODELS

Model	Accuracy	Recall	F1-score
BERT Multi-Scale Representation [22]	81	79	80
Transformer + Data Augmentation [14]	83	80	81
Hybrid RoBERTa + Linguistic Features [17]	85	84	84
Discourse + Lexical Linear Regression [23]	88	0.87	88
Large Language Models for ESL AES [24]	90	89	90
Proposed COSMET-Net	92	93	92

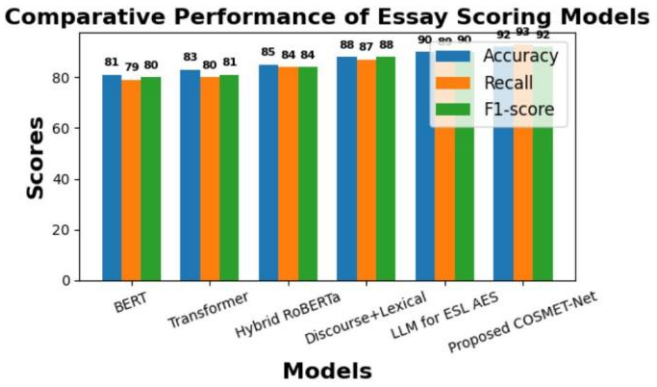

Fig. 12. Model performance comparison with percentage scores displayed.

Fig. 12 compares the performance of different essay scoring models graphically in terms of Accuracy, Recall, and F1-score, and all the values are expressed in the form of percentages. It comprises the models of BERT, Transformer-based approaches, Hybrid RoBERTa, Discourse + Lexical approaches, LLM-based approaches, and the proposed COSMET-Net. The graph also simplifies the interpretation and comparison of results between models because it displays the metrics in percentages. The COSMET-Net model is the most successful, as it has obtained the top-ranking in all the measures, which shows its strength and high score-providing capacity. Annotations are well placed, and font sizes are adjusted to make it readable and not overlap. This chart is a concise, yet informative overview of the developments in models in the area of automated essay marking.

C. Discussion

The experimental results prove that COSMET-Net provides significant gains of automated essay scoring, which is the combination of reliability, interpretability, and versatility within a single system. In addition to high performance values, the model also creates a change towards linguistically based

assessment that this type of assessment has been identified to have all along in systems of transformer-based AES. Contrastive Linguistic Regularization (CLR) allows semantic embeddings to be consistent with quantifiable linguistic features like readability ratings, lexical depth, and syntactic balance. This correspondence maximizes clarity in score rationale- a component of specific significance in educational situations that assess the interpretability, being a key factor that affects the instructional trust and usability. COSMET-Net also addresses a number of limitations noted in recent research in the AES, methodologically. Earlier studies have used contrastive learning mainly in semantic separation and have not included methods to relate embedding geometry to human-rated linguistic structures. In a similar manner, the current uses of meta learning in AES have not been combined with contrary objectives to facilitate cross-prompt adaptability in the low-resource setting. To close these methodological gaps, COSMET-Net uses the unified application of CLR, meta-learning, and contrastive loss to attain both discriminative and linguistic grounding. Such a mix is a conceptual leap to incremental architectural layering of current hybrid systems. The wider concern of these findings is that COSMET-Net does add both computational and pedagogical value. Formative assessment can be supported by the ability to produce interpretable feedback based on linguistic evidence, and the meta-learning aspect of the model will make it resistant to different topics and styles of writing. All of this makes COSMET-Net a significant step forward in the discipline that, in addition to other benefits, would provide the field with an interpretable, flexible, and linguistically aware alternative to existing transformer-based AES pipeline systems.

A contribution of interpretability that is not present in the existing literature on AES is the introduction of Contrastive Linguistic Regularization (CLR). Through imposing congruence between linguistic indicators and incorporating geometry, CLR allows formalizing an interpretive route that can determine how certain lexical, syntactic, and readability cues can be used to affect scoring decisions. This is unlike other transformer-based models that provide attention maps without basing them on quantifiable linguistic constructs. The feedback explicable by this correspondence has a pedagogical meaning, as the identified linguistic features are associated with the criteria applied by human teachers, thus helping to facilitate formative learning and self-directed improvement. In addition, its meta-learning aspect increases resilience in real-world application scenarios due to quick adaptation to new prompts and situations of writing. This property decreases the sensitivity of AES systems to prompt-specific distributions, a well-known weakness in practice in educational testing settings because of the topic variety and lack of data, which are commonly experienced. Taken together, these contributions put COSMET-Net to the forefront as a performance-based system as well as an interpretable, pedagogically useful, and operationally sound academic writing evaluation framework.

D. Comparative Advantages Over Existing AES Models

COSMET-Net brings a number of benefits in comparison with the methods of automated essay scoring and the associated hybrid architecture. The main assumption of traditional transformer-based models deals with the quality of semantic representations, but lacks explicit methods of grounding predictions based on linguistic indicators by human evaluators. Conversely, the Contrastive Linguistic Regularization (CLR) aspect of COSMET-Net is designed to guarantee that embedding structures incorporate linguistic interpretability in the form of readability, syntactic variation and lexical richness. This makes the rationale of scoring more transparent than attention-based interpretability methods. Models that use manually crafted linguistic representations tend to be less rich in semantic insight than the transformer embeddings, where contrastive learning methods in previous AES experiments tend to optimize semantic distinctness without relating these representations to linguistic concepts. COSMET-Net seals this with a hybrid discriminative power of contrastive learning alongside the use of linguistically based regularization. Furthermore, the meta-learning feature enables flexibility that is not common to AES pipelines, and the model will be able to maintain a high level of performance when subjected to cross-prompt and low-resource conditions. All these benefits bring COSMET-Net as more of an interpretable, adaptable, and pedagogically suited approach to the current practices in the field.

V. Conclusion and Future Work

In this study, a united framework, COSMET-Net, aimed at improving automated essay grading by combining contrastive learning, meta-learning adaptation, and Contrastive Linguistic Regularization (CLR) was proposed. The experimental outcome showed that the framework not only meets the competitive performance in terms of accuracy, recall, and F1-score, but also provides interpretability benefits by matching the embedding structure with quantifiable linguistic measures. Such a combination of semantic accuracy, linguistic foundation, and cross-prompt flexibility reflects a multidimensional improvement in AES skill. The reflections of the experiments disclose three main contributions: The CLR mechanism enhances the interpretive connection between deep representations and linguistic features applied in human scoring; the meta-learning component enhances sensitivity to prompt variation, allowing the use of competent performance in low-resource or unseen conditions; and the unified design facilitates equal gains on predictive accuracy, transparency, and pedagogical relevance. These empirical results highlight the practical and theoretical importance of the method. The wider implications of the use of performance metrics are in educational technology, where we need explainable assessment and cross-context generalization in order to have trustworthy and scalable use. COSMET-Net goes a step further to provide an end-to-end AES architecture that can concurrently mitigate gaps in interpretability, provide enhanced adaptability, and align algorithmic output to human-assessed linguistic constructs. This contribution makes the framework a significant contribution to the research in automated writing assessment. Future research could investigate the possibility of extending the model into other types of writing, multilingual scoring, and real-time feedback application to increase the pedagogical role of the model.

The next step in the development of COSMET-Net will be the extension of the capabilities to include more writing genres and academic fields. It can be further enhanced with domain-specific language models and adaptive prompt-tuning strategies,

particularly in the specialized setting. Also, it would be advisable to introduce real-time feedback systems and interactive dashboards to enable users to get immediate recommendations on how to improve. The investigation of multi-modes such as that of text and spoken or visual stimuli might be used to improve learning. The solutions to the issues linked with equity, the reduction of bias, and cultural diversity in essay collections will be essential in the context of wider adoption. Furthermore, it is possible to make the framework extend to low-resource languages and offline environments so that every learner can be accessible. Finally, the research in the future will attempt to transform COSMET-Net into not a strong performing scoring tool but also a smart writing assistant that will encourage self-directed learning and comprehensive academic growth.

REFERENCES

[1] S. Rawas, "ChatGPT: Empowering lifelong learning in the digital age of higher education," Education and Information Technologies, vol. 29, no. 6, pp. 6895–6908, 2024.

[2] M. Omar, S. Choi, D. Nyang, and D. Mohaisen, "Robust natural language processing: Recent advances, challenges, and future directions," IEEE Access, vol. 10, pp. 86038–86056, 2022.

[3] J. Y. Bai et al., "Automated essay scoring (AES) systems: Opportunities and challenges for open and distance education," in Proceedings of The Tenth Pan-Commonwealth Forum on Open Learning (PCF10), 2022.

[4] B. S. Latibari et al., "Transformers: A security perspective," IEEE Access, 2024.

[5] M. U. Hadi et al., "Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects," Authorea preprints, vol. 1, no. 3, pp. 1–26, 2023.

[6] V. Hassija et al., "Interpreting black-box models: a review on explainable artificial intelligence," Cognitive Computation, vol. 16, no. 1, pp. 45–74, 2024.

[7] S. Li et al., "Vitatecs: A diagnostic dataset for temporal concept understanding of video-language models," in European Conference on Computer Vision, Springer, 2024, pp. 331–348.

[8] E. Hashmi, S. Y. Yayilgan, and M. Abomhara, "Metalinguist: enhancing hate speech detection with cross-lingual meta-learning," Complex & Intelligent Systems, vol. 11, no. 4, p. 179, 2025.

[9] S. Sarabi Asl, M. Rashtchi, and G. Rezaie, "The effects of interactionist versus interventionist dynamic assessment models on Iranian EFL learners' speaking sub-skills: a mixed-method study," Asian-Pacific Journal of Second and Foreign Language Education, vol. 9, no. 1, p. 12, 2024.

[10] D. Zaikis and I. Vlahavas, "From pre-training to meta-learning: a journey in low-resource-language representation learning," IEEE Access, vol. 11, pp. 115951–115967, 2023.

[11] D. Soydaner and J. Wagemans, "Unveiling the factors of aesthetic preferences with explainable AI," British Journal of Psychology, 2024.

[12] L. J. Jacobsen and K. E. Weber, "The promises and pitfalls of large language models as feedback providers: A study of prompt engineering and the quality of AI-driven feedback," AI, vol. 6, no. 2, p. 35, 2025.

[13] G. Palma, G. Cecchi, M. Caronna, and A. Rizzo, "Leveraging Large Language Models for Scalable and Explainable Cybersecurity Log Analysis," Journal of Cybersecurity and Privacy, vol. 5, no. 3, p. 55, 2025.

[14] K. Gupta, "Data augmentation for automated essay scoring using transformer models," in 2023 International Conference on Artificial Intelligence and Smart Communication (AISC), IEEE, 2023, pp. 853–857.

[15] Y. Song, Q. Zhu, H. Wang, and Q. Zheng, "Automated essay scoring and revising based on open-source large language models," IEEE Transactions on Learning Technologies, vol. 17, pp. 1880–1890, 2024.

[16] X. Tang, H. Chen, D. Lin, and K. Li, "Incorporating fine-grained linguistic features and explainable ai into multi-dimensional automated writing assessment," Applied Sciences, vol. 14, no. 10, p. 4182, 2024.

[17] M. Faseeh et al., "Hybrid approach to automated essay scoring: Integrating deep learning embeddings with handcrafted linguistic features for improved accuracy," Mathematics, vol. 12, no. 21, p. 3416, 2024.

[18] F. Li, X. Xi, Z. Cui, D. Li, and W. Zeng, "Automatic essay scoring method based on multi-scale features," Applied Sciences, vol. 13, no. 11, p. 6775, 2023.

[19] T. Amin, F. Aadil, K. M. Awan, S. Lim, and others, "Enhancing Essay Scoring: An Analytical and Holistic Approach With Few-Shot Transformer-Based Models," IEEE Access, 2025.

[20] A. Pack, A. Barrett, and J. Escalante, "Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability," Computers and Education: Artificial Intelligence, vol. 6, p. 100234, June 2024, doi: 10.1016/j.caeai.2024.100234.

[21] Ibrahimmazlum, "IELTS Writing Scored Essays Dataset." [Online]. Available: https://www.kaggle.com/datasets/mazlumi/ielts-writing-scored-essays-dataset

[22] Y. Wang, C. Wang, R. Li, and H. Lin, "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, M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, Eds., Seattle, United States: Association for Computational Linguistics, July 2022, pp. 3416–3425. doi: 10.18653/v1/2022.naacl-main.249.

[23] H. M. Alawadh, T. Meraj, L. Aldosari, and H. Tayyab Rauf, "An Efficient Text-Mining Framework of Automatic Essay Grading Using Discourse Macrostructural and Statistical Lexical Features," SAGE Open, vol. 14, no. 4, p. 21582440241300548, Oct. 2024, doi: 10.1177/21582440241300548.

[24] A. Pack, A. Barrett, and J. Escalante, "Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability," Computers and Education: Artificial Intelligence, vol. 6, p. 100234, June 2024, doi: 10.1016/j.caeai.2024.100234.