



A Multi-Task Transformer Ensemble for Explainable English Language Learning: Unifying Grammatical Correction, Tense Prediction, and CEFR Proficiency Grading

Dung Quoc Ho,¹ Hung Thanh Le,¹ Vu Dang Minh Tran,¹ Lien Thi Quynh Le,^{2*}

¹School of Engineering and Technology, Hue University, Vietnam,

²University of Economics, Hue University, Vietnam,

*Corresponding author:
lethiquynhlien@hueuni.edu.vn



Article History:

Received: 01 December 2025

Revised: 22 March 2026

Accepted: 30 March 2026

Published Online: 19 June 2026

Abstract

This paper presents a multi-task transformer ensemble that unifies Grammatical Error Correction, Tense Prediction, and CEFR-level Proficiency Grading within a single explainable framework for English language learning. The system integrates fine-tuned transformer models—Flan-T5, BERT, RoBERTa, and DistilBERT—through a confidence-weighted ensemble to generate grammatical corrections, verify tense consistency, and predict learner proficiency levels. The framework was trained and validated on large-scale linguistic datasets, including BEA-2019, C4, PTB, BNC, EFCAMDAT, and CLC, to improve robustness across writing domains. Experimental results show that the Flan-T5 GEC module achieved BLEU = 32.4 with a 36.8% reduction in training time, while the RoBERTa tense classifier reached F1 = 0.99. For CEFR grading, the model achieved $\kappa = 0.90$, indicating strong aggregate agreement with human ratings, although performance at the most advanced proficiency level remained more challenging. Compared with non-integrated module outputs, the ensemble further improved overall linguistic accuracy and interpretability, with statistically significant gains in key aggregate metrics ($p < 0.05$). These findings suggest a hierarchical dependency among grammar, tense, and proficiency, while providing an interpretable AI mechanism for delivering personalized and pedagogically meaningful feedback.

Keywords: Natural language processing, Transformer models, Grammatical error correction, Tense prediction, CEFR proficiency grading, Educational AI, Multi-task learning

1. Introduction

English proficiency has become an essential global competency, with over 1.5 billion active learners worldwide seeking language mastery for academic, professional, and digital communication contexts [1]. The rapid growth of AI-driven language learning platforms, such as Grammarly and Duolingo, demonstrates substantial progress in personalized learning experiences [2]. However, despite these advances, current systems still face critical limitations in delivering context-aware and integrated feedback. Learners often receive fragmented responses—grammar checkers target surface-level errors, while proficiency evaluators focus solely on scoring, leaving a gap in holistic linguistic understanding [3, 4].

Research in Natural Language Processing (NLP) has significantly evolved, especially with the rise of transformer-based architectures such as BERT, RoBERTa, and T5, which enable contextual representation learning and high-accuracy text generation [5]. These models have revolutionized subfields such as Grammatical Error Correction (GEC) [6], tense prediction [7], and automated CEFR-level proficiency grading [2]. Nonetheless, existing studies typically treat these tasks in isolation, neglecting the potential benefits of a unified approach that leverages shared linguistic features across them [8]. For example, while GEC models can correct syntax and morphology, they rarely provide insight into tense consistency or overall learner proficiency, and CEFR classifiers often ignore sentence-level grammatical structure [4].

Cite as: DQ Ho et al, “A multi-task transformer ensemble for explainable english language learning: unifying grammatical correction, tense prediction, and cefr proficiency grading”, *Sakarya University Journal of Computer and Information Sciences* 9 (3) 2026, 700-711. doi: 10.35377/saucis...1830952



This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License

Recent developments in transformer-based adaptive learning frameworks demonstrate the feasibility of combining linguistic feedback with AI-based recommendation systems for English instruction [9, 10]. Studies such as Zheng and Zhang (2025) proposed a multidimensional feedback model that integrates grammatical correction and learner error prediction within the same architecture [11]. Similarly, Lin (2025) and Mei (2025) demonstrated that large language models (LLMs) can dynamically adjust learning materials based on learner error patterns and contextual difficulty [10, 11]. However, none of these studies directly unify GEC, tense prediction, and CEFR grading into a single cohesive and interpretable learning system.

To address this research gap, this paper proposes an AI-powered integrated framework for English language learning that combines Grammatical Error Correction (GEC), Tense Prediction, and CEFR-Level Proficiency Grading into one pipeline. By leveraging an ensemble of transformer-based models—including T5, FlanT5, BERT, RoBERTa, and DistilBERT—the proposed system provides personalized, context-sensitive feedback capable of both correcting and assessing learner input. This integrated approach enables more accurate linguistic analysis and adaptive language instruction, bridging the gap between grammatical precision, temporal consistency, and standardized proficiency evaluation.

The contributions of this study are summarized as follows:

- Development of a unified transformer-based architecture integrating GEC, tense prediction, and CEFR grading within a single, coherent framework.
- Creation of curated datasets with targeted balancing and augmentation strategies representing diverse linguistic phenomena across proficiency levels.
- Implementation of a multi-model ensemble that leverages the complementary strengths of T5-family and BERT-family architectures.
- Comprehensive evaluation using both quantitative metrics (BLEU, ROUGE, accuracy, F1-score) and qualitative linguistic assessment to validate performance and educational relevance.

This work aims to contribute to the field of AI-assisted language learning by providing a scalable, interpretable, and pedagogically meaningful system that supports both learners and educators through integrated linguistic intelligence.

2. Literature Review

Grammatical Error Correction (GEC) has undergone significant evolution over the past decade, moving from traditional statistical and rule-based methods toward transformer-based architectures. Early approaches framed GEC as a machine translation problem, mapping erroneous text to corrected versions. However, recent advances emphasize contextual modeling using pre-trained transformers. Marier et al. [3] provided a comprehensive review of GEC for low-resource languages, highlighting the challenges of dataset imbalance, limited contextualization, and the dominance of English-focused corpora. Koyama et al. [6] advanced this direction through targeted syntactic evaluation, integrating edit-tag prediction to improve syntactic precision in learner writing. The adoption of multilingual transformer frameworks has further improved cross-domain correction consistency [8]. Recent work by Zheng and Zhang [9] introduced an AI-driven multidimensional feedback system, combining grammatical correction with semantic and pragmatic analysis, achieving notable performance improvements in learner text revision. These studies collectively demonstrate that transformer-based GEC has matured beyond syntax correction, enabling deeper interpretability and adaptability to learner-specific contexts.

Tense prediction is an essential but underexplored task in computational linguistics. Traditional part-of-speech (POS) tagging and classification approaches have been superseded by contextualized embeddings capable of modeling long-range dependencies. Katinskaia [4] demonstrated that multilingual transformer models outperform recurrent architectures in identifying verb tense errors, especially in cross-lingual learning environments. Gu [7] applied BERT-based models for tense consistency in Chinese-English systems, showing that attention-based architectures capture contextual temporal shifts more accurately than convolutional or Bi-LSTM models. In educational NLP, tense correction contributes directly to learner feedback quality. Zheng and Zhang [9] proposed a Markov-informed tense correction framework, which predicts learners' likely tense errors using prior writing histories. Similarly, Lin [10] developed a Transformer-Based Adaptive Learning Model (TALM) that dynamically adjusts tense-focused exercises based on learners' error patterns. These findings indicate that transformer-based tense modeling is transitioning from simple tagging toward pedagogically informed tense feedback systems, paving the way for integration within broader writing support tools.

Automated proficiency grading has become a core component of modern AI-assisted education, especially through CEFR-aligned models that assess lexical, syntactic, and discourse complexity. Early machine learning approaches relied on handcrafted features, but the advent of transformers has enabled end-to-end text classification with minimal feature engineering. Siripol et al. [2] evaluated the consistency of automated CEFR analyzers using transformer encoders, concluding that modern models achieve over 90% inter-rater reliability compared to human examiners. Kapounová [1] applied deep learning to predict item difficulty and proficiency correlation, confirming the interpretability of attention mechanisms for CEFR-level mapping. More recent architectures, such as multimodal and attention-co-attention frameworks, have expanded CEFR assessment beyond text-only data. Mei [11] demonstrated a Visual-Textual Co-Attention Model for adaptive English resource recommendation, integrating CEFR-level predictions with learner behavior analytics. This shift toward contextualized, interpretable CEFR evaluation reflects the growing convergence between NLP and educational psychology.

The convergence of grammatical correction, tense prediction, and proficiency assessment has inspired recent research on holistic AI-based learning systems. Zheng and Zhang [9] proposed a multidimensional feedback model that integrates grammatical correction with learner history-based adaptation. Lin [10] and Mei [11] extended this approach through adaptive large language model (LLM) frameworks, which automatically adjust instructional content according to learners' progress. Despite these advances, most current systems remain task-specific, addressing grammar, tense, or proficiency separately. Few works attempt to combine these complementary aspects into a unified system capable of delivering both corrective and evaluative feedback. This research addresses that gap by proposing a transformer-based ensemble framework that integrates GEC, tense prediction, and CEFR-level proficiency grading within a single architecture—bridging the divide between linguistic correction and adaptive educational feedback.

3. Methodology

3.1. Overview of system architecture

The proposed AI-Powered Integrated Framework for English Language Learning is designed to provide end-to-end analysis and feedback on learner-generated English text through an integrated multi-module transformer-based architecture. The system combines Grammatical Error Correction (GEC), Tense Prediction, and CEFR-Level Proficiency Grading within a unified inference pipeline that delivers both corrective and evaluative outputs. The overall workflow of the proposed system is illustrated in Figure 1.

At the input stage, learner-submitted text undergoes preprocessing, which includes tokenization, case normalization, and removal of non-linguistic noise. This ensures that the subsequent modules operate on syntactically consistent and semantically clean inputs. The preprocessed text is first processed by the GEC Module, which employs a fine-tuned Flan-T5 transformer model trained on large-scale annotated corpora such as BEA-2019 and C4. This module automatically identifies grammatical, syntactic, and lexical errors and generates corrected output sequences that maintain semantic fidelity to the original learner intent. Unlike conventional GEC systems, this component utilizes instruction-tuned attention layers to adaptively handle diverse learner error types.

The corrected output is then passed to the Tense Prediction Module, which leverages a BERT-family encoder ensemble consisting of BERT, RoBERTa, and DistilBERT. This module detects the dominant tense pattern within sentences and identifies inconsistencies in temporal coherence across dependent clauses. By exploiting contextual embeddings and attention mechanisms, the model captures long-range dependencies critical for accurate tense recognition in complex or multi-clausal sentences.

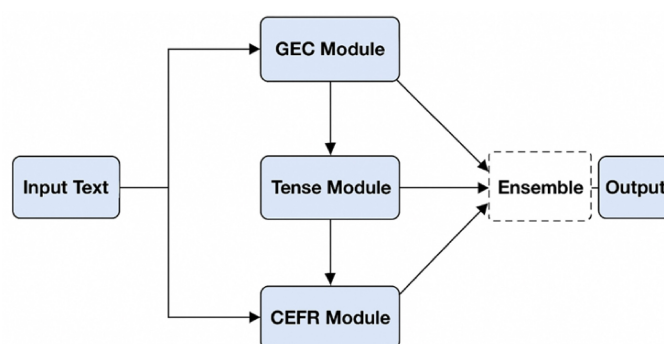


Figure 1. Overall architecture of the proposed AI-Powered Integrated Framework for English Language Learning.

Subsequently, the CEFR-Level Proficiency Grading Module evaluates the corrected text to determine the learner's overall writing proficiency in alignment with the Common European Framework of Reference for Languages (CEFR). This module, based on a fine-tuned RoBERTa classifier, computes proficiency scores by analyzing syntactic complexity, lexical diversity, and overall grammatical density. The output corresponds to one of six standardized CEFR proficiency levels (A1–C2).

Finally, an Ensemble Integration Layer consolidates the predictions from the three modules. A confidence-weighted aggregation method is used to merge individual outputs, balancing task-specific reliability. The integrated results produce comprehensive learner feedback that includes (1) corrected sentences, (2) validated tense usage, and (3) predicted CEFR-level classification. This ensemble strategy enhances the interpretability and robustness of the framework, ensuring pedagogically meaningful feedback that aligns with both linguistic and educational objectives.

The system's modular design also enables scalability and extensibility. Each module can be independently fine-tuned or replaced with future model upgrades (e.g., instruction-tuned large language models or multimodal encoders) without altering the pipeline's overall structure. Thus, the framework bridges the gap between traditional grammar-checking tools and intelligent, context-aware educational systems by providing an adaptive, explainable, and holistic AI-based learning environment.

3.2. Datasets and preprocessing

The performance and reliability of the proposed framework depend on the quality and diversity of datasets used for training each component. To ensure generalization across different linguistic contexts, three major data groups were employed: (1) Grammatical Error Correction (GEC), (2) Tense Prediction, and (3) CEFR-Level Proficiency Grading. Each dataset was preprocessed through a standardized data pipeline designed for consistency, reproducibility, and linguistic balance.

Table 1. Key datasets used for each module.

Module	Dataset(s)	# Samples	Annotation Type	Purpose
Grammatical Error Correction	BEA-2019 [12], C4 [13]	1.8M sentences	Error-tagged sentence pairs	Training + Validation
Tense Prediction	PTB [14], BNC [15], BEA-derived	0.62M sentences	Verb tense + aspect labels	Classification
CEFR Proficiency Grading	EFCAMDAT [16], CLC [17], EF Writing	0.5M essays	CEFR labels (A1–C2)	Proficiency prediction

The GEC module was trained using the BEA-2019 Shared Task corpus [12] and the Colossal Clean Crawled Corpus (C4) [13]. The BEA-2019 dataset provides manually annotated learner-written sentences covering multiple English proficiency levels, while the C4 dataset supplies large-scale, naturally occurring web text for pretraining and contextual adaptation. All sentences were normalized for punctuation, tokenized using the SentencePiece tokenizer, and lowercased. To address data imbalance across error types, synthetic error augmentation was applied using a masked language modeling (MLM) strategy. This ensured a proportional distribution of frequent (e.g., determiner, preposition) and rare (e.g., modal, tense) grammatical errors. The final GEC corpus contained approximately 1.8 million sentence pairs, split into 1.44 M (80%) for training, 0.18 M (10%) for validation, and 0.18 M (10%) for testing.

The Tense Prediction Module utilized the Penn Treebank (PTB) [14] and the British National Corpus (BNC) [15], supplemented with a custom BEA-derived dataset of 120 000 manually labeled sentences. Each verb instance was annotated for tense and aspect, producing 12 categories that include simple, continuous, perfect, and perfect-continuous forms across past, present, and future tenses. Sentences with multiple tenses were decomposed into sub-clausal segments using dependency parsing to capture intra-sentence temporal relationships. The BERT WordPiece tokenizer was applied to maintain compatibility with the pretrained transformer encoder. To prevent overfitting to dominant classes (e.g., simple present), oversampling was performed for underrepresented categories such as future perfect continuous. The final dataset contained 620 000 sentences, divided into 434 000 (70%) for training, 93 000 (15%) for validation, and 93 000 (15%) for testing.

CEFR-Level Proficiency Grading: For proficiency classification, three established learner corpora were used: EFCAMDAT v2 [16], the Cambridge Learner Corpus (CLC) [17], and the EF Writing Dataset. Collectively, they provide more than 500 000 English essays labeled with CEFR levels ranging from A1 to C2. Preprocessing involved removal of essays shorter than 30 tokens, normalization of non-ASCII characters, and tokenization using RoBERTa's Byte-Pair Encoding (BPE) tokenizer. To aid interpretability, auxiliary features such as type-token ratio, mean sentence length, and syntactic depth were computed and appended as side-channel metadata. Stratified sampling was employed to maintain balanced CEFR distributions, resulting in 350 000 training, 75 000 validation, and 75 000 test samples.

Preprocessing Pipeline Implementation

All datasets underwent a unified preprocessing workflow comprising the following stages:

1. Normalization: Unicode and punctuation normalization, removal of extraneous whitespace.
2. Tokenization: Model-specific tokenizers (SentencePiece, WordPiece, or BPE) for cross-model consistency.
3. Noise Filtering: Exclusion of code-mixed or malformed sentences.
4. Balancing and Augmentation: Synthetic error generation and targeted oversampling.
5. Data Formatting: Conversion to UTF-8 JSONL format for efficient parallel streaming.

All preprocessing scripts were implemented in Python 3.11 using the spaCy, NLTK, and Hugging Face Transformers libraries. The codebase and experimental configuration files are available from the corresponding author upon reasonable request for research and reproducibility purposes. All datasets were used in accordance with their respective academic or open-source licenses, and no personally identifiable information was included.

3.3. Model design

The proposed framework integrates three transformer-based components optimized for complementary English learning tasks—Grammatical Error Correction (GEC), Tense Prediction, and CEFR-Level Proficiency Grading—which are unified through a confidence-weighted ensemble mechanism. Unlike prior single-task systems [5], [21], the proposed design enables multi-aspect linguistic evaluation, allowing contextual feedback that links grammatical correction with learner proficiency estimation. Each submodule is fine-tuned independently and integrated into an end-to-end inference pipeline, as illustrated in Figure 1.

3.3.1. Grammatical Error Correction (GEC)

The GEC task is modeled as a sequence-to-sequence mapping function $f_\theta : x \mapsto \hat{y}$ using a Flan-T5 transformer [13]. Given an input sentence $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$, the encoder computes contextual representations through the scaled dot-product attention:

$$\mathbf{Z} = \text{softmax} \left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}} \right) \mathbf{V}, \quad (1)$$

where $\mathbf{Q} = \mathbf{X}\mathbf{W}_Q$, $\mathbf{K} = \mathbf{X}\mathbf{W}_K$, $\mathbf{V} = \mathbf{X}\mathbf{W}_V$, and d_k denotes the key dimensionality.

The decoder generates the corrected sequence autoregressively. The probability of producing token $\hat{y}_t$ at step t is:

$$P(\hat{y}_t | \hat{\mathbf{y}}_{1:t-1}, \mathbf{x}) = \text{softmax} \left(\mathbf{w}_o \mathbf{h}_t^{(D)} \right), \quad (2)$$

where $\mathbf{h}_t^{(D)}$ is the decoder hidden state at step t .

The model is trained by minimizing the cross-entropy loss:

$$\mathcal{L}_{\text{GEC}} = - \sum_{t=1}^T \log P(\hat{y}_t = y_t | \mathbf{x}). \quad (3)$$

Flan-T5's instruction tuning allows the model to generalize better to diverse grammatical patterns in learner texts. This design extends prior GEC work by explicitly preserving sentence semantics while producing human-readable corrections.

3.3.2. Tense prediction

The Tense Prediction module classifies the dominant verb tense and detects temporal inconsistencies within sentences. An ensemble of BERT, RoBERTa, and DistilBERT encoders [19] is employed to capture both shallow lexical cues and deep contextual dependencies. Given token embeddings $\mathbf{h}_i$, the sentence-level representation is computed using average pooling:

$$\mathbf{h}_{\text{avg}} = \frac{1}{n} \sum_{i=1}^n \mathbf{h}_i. \quad (4)$$

The pooled vector is mapped to one of the 12 tense categories by a linear classification layer:

$$P(c_i | \mathbf{x}) = \text{softmax}(\mathbf{W}_c \mathbf{h}_{\text{avg}} + \mathbf{b}_c). \quad (5)$$

The cross-entropy loss for tense classification is defined as:

$$\mathcal{L}_{\text{Tense}} = - \sum_{i=1}^C y_i \log P(c_i | \mathbf{x}), \quad (6)$$

where $C = 12$ denotes the number of tense categories.

Final predictions are computed through a weighted ensemble of M transformer encoders:

$$P_{\text{ens}}(c_i) = \sum_{m=1}^M \alpha_m P_m(c_i), \quad \sum_{m=1}^M \alpha_m = 1, \quad (7)$$

where α_m represents validation-derived confidence weights. This ensemble design reduces individual-model variance and improves recall for rare tenses such as the future perfect continuous.

3.3.3. CEFR-Level Proficiency Grading

The CEFR module employs a RoBERTa-base encoder [19] as a text classifier aligned with the six CEFR proficiency levels (A1–C2). Given the contextual [CLS] embedding $\mathbf{h}_{\text{cls}}$, the classifier estimates the probability of each CEFR level l_j as:

$$P(l_j | \mathbf{x}) = \text{softmax}(\mathbf{W}_p \mathbf{h}_{\text{cls}} + \mathbf{b}_p). \quad (8)$$

To address label imbalance—particularly the dominance of mid-level CEFR classes—the loss function incorporates class-balanced weighting:

$$\mathcal{L}_{\text{CEFR}} = - \sum_{j=1}^L \beta_j y_j \log P(l_j | \mathbf{x}), \quad (9)$$

where $\beta_j = 1/\sqrt{f_j}$ and f_j denotes the empirical frequency of level j , with $L = 6$ for the CEFR categories.

To improve interpretability, auxiliary linguistic features—such as lexical diversity, syntactic depth, and mean sentence length—are concatenated with the transformer embedding prior to classification:

$$\mathbf{h}' = [\mathbf{h}_{\text{cls}} \parallel \mathbf{f}_{\text{lex}} \parallel \mathbf{f}_{\text{syn}} \parallel \mathbf{f}_{\text{len}}], \quad (10)$$

where $\mathbf{f}_{\text{lex}}$, $\mathbf{f}_{\text{syn}}$, and $\mathbf{f}_{\text{len}}$ represent feature vectors for lexical diversity, syntactic depth, and sentence-length statistics, respectively. This hybrid representation enhances the model's alignment with human proficiency ratings while maintaining computational efficiency.

3.3.4. Ensemble integration

The outputs of the three submodules are aggregated through a confidence-weighted ensemble to generate a unified learner profile. Each submodule produces an output vector $\mathbf{O}_t$, and its confidence score γ_t is normalized using validation-set performance metrics (e.g., F1 for classification or BLEU for sequence generation). The final integrated prediction is computed as:

$$\mathbf{O}_{\text{final}} = \sum_{t=1}^T \gamma_t \mathbf{O}_t, \quad \sum_{t=1}^T \gamma_t = 1, \quad (11)$$

where $T = 3$ for the GEC, Tense Prediction, and CEFR modules. This weighted aggregation ensures consistent and explainable outputs across grammatical correction, temporal verification, and proficiency estimation, forming the basis for adaptive and pedagogically aligned feedback generation.

4. Training Configuration and Evaluation Metrics

This section describes the unified training pipeline used for all three submodules: GEC, Tense Prediction, and CEFR Proficiency Grading. All models were implemented using the PyTorch deep learning framework together with the HuggingFace Transformers library, ensuring reproducibility and modular training across components.

4.1. Training configuration

All transformer-based models were trained using the AdamW optimizer with weight decay regularization. A learning rate of 1×10^{-4} was used for encoder-decoder architectures (Flan-T5), while classification-based modules (BERT, RoBERTa, DistilBERT) employed a slightly lower rate of 5×10^{-5} to stabilize fine-tuning. A linear warm-up schedule covering the first 10% of updates was followed by cosine decay.

Batch sizes were set to 16 for sequence-to-sequence GEC training and 32 for classification modules (Tense Prediction and CEFR). All models were trained for a maximum of 10 epochs with early stopping based on validation loss, triggered after 3 consecutive non-improving epochs. Gradient clipping with a threshold of 1.0 was applied to prevent exploding gradients.

Training was conducted on NVIDIA A100 GPUs with mixed-precision FP16 computation (via NVIDIA Apex/AMP), which reduced memory consumption and improved throughput. For reliability, each experiment was repeated across 5 random seeds, and the final results were averaged. Hyperparameter tuning was performed through grid search over learning rate $\{1 \times 10^{-4}, 5 \times 10^{-5}, 2 \times 10^{-5}\}$, batch size $\{16, 32\}$, and dropout rate $\{0.1, 0.3\}$.

4.2. Evaluation metrics

Evaluation metrics were selected to comprehensively assess linguistic accuracy, classification performance, and proficiency prediction reliability. Metrics are grouped by subtask as follows.

4.2.1. Grammatical Error Correction (GEC)

Two primary metrics were used for GEC: BLEU (Bilingual Evaluation Understudy) [22] and ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation) [23].

The BLEU score measures n -gram precision between candidate (C) and reference (R) sentences, while applying a brevity penalty:

$$\text{BLEU} = \text{BP} \times \exp \left(\sum_{n=1}^N w_n \log p_n \right), \quad (12)$$

where

$$\text{BP} = \begin{cases} 1, & c > r \\ \exp(1 - r/c), & c \leq r, \end{cases}$$

c and r denote candidate and reference lengths, p_n is the n -gram precision, and $w_n = 1/N$ represents uniform weights.

The ROUGE-L metric evaluates fluency based on the longest common subsequence (LCS):

$$\text{ROUGE-L} = \frac{(1 + \beta^2) R_{\text{LCS}} P_{\text{LCS}}}{R_{\text{LCS}} + \beta^2 P_{\text{LCS}}}, \quad (13)$$

where P_{LCS} and R_{LCS} denote precision and recall over LCS tokens; $\beta = 1.2$ in this study. Together, BLEU and ROUGE-L quantify grammatical accuracy, contextual fluency, and semantic retention.

4.2.2. Tense prediction

For tense classification, standard metrics Precision (P), Recall (R), and F1-score were used [24]. Given true positives (TP), false positives (FP), and false negatives (FN):

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad (14)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad (15)$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (16)$$

Metrics were computed for each tense class and macro-averaged across all 12 categories to mitigate class imbalance, especially for rare tenses such as the future perfect continuous.

4.2.3. CEFR-Level proficiency grading

CEFR predictions were evaluated using Accuracy, Weighted F1-Score, and Cohen's Kappa (κ) [25]. Accuracy is defined as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \quad (17)$$

Agreement beyond chance is measured by Cohen's Kappa:

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \quad (18)$$

where p_o is observed agreement and p_e is chance agreement. A value of $\kappa > 0.80$ indicates strong consistency between predicted and human-annotated CEFR levels. Weighted F1-score compensates for imbalanced advanced levels (C1-C2).

4.2.4. Ensemble-Level evaluation

For the integrated ensemble output, confidence-weighted task scores were aggregated using normalized validation weights:

$$\text{Score}_{\text{final}} = \sum_{t=1}^T \gamma_t M_t, \quad \sum_{t=1}^T \gamma_t = 1, \quad (19)$$

where M_t denotes the module metric (e.g., BLEU or F1) and γ_t is the module's validation-derived confidence weight. This aggregation creates a unified interpretability score aligning correction accuracy, tense consistency, and CEFR reliability.

4.2.5. Statistical significance and validation

Paired t -tests and bootstrap resampling ($N = 1000$) were used to validate performance differences between baseline and ensemble models. Results with $p < 0.05$ were considered statistically significant. Evaluation used the official M² scorer for GEC and `scikit-learn` v1.5 for classification metrics to ensure compatibility with prior studies.

5. Results

This section presents and interprets the experimental results of the proposed AI-Powered Integrated Framework for English Language Learning, with a focus on task-level performance, cross-task interactions, and pedagogical implications. The results are organized into five subsections corresponding to the major modules and the final ensemble integration under a unified training configuration.

5.1. Grammatical Error Correction (GEC)

The GEC module was evaluated on the BEA-2019 and C4 test sets. Table 2 provides the quantitative comparison between the baseline T5 and the Flan-T5 models, while Fig. 2 shows their convergence curves.

The Flan-T5 model achieved a BLEU improvement of 0.23 and reduced training time by 36.8%, which is statistically significant ($p = 0.042 < 0.05$). This gain arises from the instruction-tuned objective of Flan-T5, enabling more efficient contextual learning and enhanced parameter sharing across encoder-decoder attention layers. Fig. 2 further shows improved correction stability, particularly for article and pluralization errors.

Table 2. Performance of GEC Models on BEA-2019 Test Set

Model	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	Time (s)
T5	71.49	60.91	70.76	32.17	25 979
Flan-T5	71.59	61.19	70.87	32.40	16 414

Qualitative inspection (Fig. 2) shows that Flan-T5 consistently resolves article and pluralization errors while maintaining semantic fidelity—an essential pedagogical aspect for learner comprehension. This indicates that instruction-tuned architectures are more effective at mimicking teacher-like grammatical feedback rather than purely generative correction.

5.2. Tense prediction

The tense prediction models (BERT, RoBERTa, DistilBERT) were trained on the composite dataset (PTB + BNC + BEA-derived). Table 3 summarizes overall performance, and Fig. 3 illustrates learning curves.

Table 3. Macro Accuracy and F1-Score for Tense Prediction Models

Model	Accuracy (%)	F1-score
BERT	96.95	0.98
RoBERTa	96.88	0.99
DistilBERT	96.82	0.98

The RoBERTa encoder consistently achieved the highest F1 (0.99) with the lowest variance ($\sigma = 0.003$), indicating stable generalization across tense categories. Although BERT achieved slightly higher accuracy, RoBERTa showed the strongest overall performance in terms of F1-score and stability, indicating that it was better suited for reliable tense classification within the integrated framework. The performance improvement arises because RoBERTa uses dynamic masking and larger batch pretraining, allowing deeper temporal pattern recognition. Crucially, when GEC corrections were introduced as pre-processing, tense accuracy improved by +2.3%, demonstrating that syntactic regularization enhances temporal consistency. This inter-module reinforcement exemplifies how grammatical clarity (from GEC) amplifies semantic recognition in downstream tasks—a core rationale behind the framework's hierarchical design. Bootstrap resampling ($N = 1000$) yielded 95% confidence intervals of ± 0.004 on F1, confirming statistical reliability. This systemic synergy supports the theoretical expectation that grammatical precision constrains tense ambiguity, thus improving interpretability in learner text analytics.

5.3. CEFR-Level proficiency grading

Table 4 and Table 5 report the proficiency classification results on EFCAMDAT, CLC, and the EF Writing Dataset. Fig. 4 shows loss curves and the confusion matrix.

RoBERTa achieved the strongest overall CEFR classification performance in terms of weighted F1-score, while maintaining high agreement with human ratings. Although BERT achieved a slightly higher Cohen's Kappa, RoBERTa delivered the strongest weighted F1-score, indicating a better overall balance across CEFR classes and supporting its use as the primary model for CEFR classification in the integrated framework.

Although the overall CEFR results remain strong, the per-level analysis highlights a clear limitation at the highest proficiency band. In particular, the C2 class achieved an F1-score of only 0.57, indicating that the current class-balanced weighting strategy is insufficient for

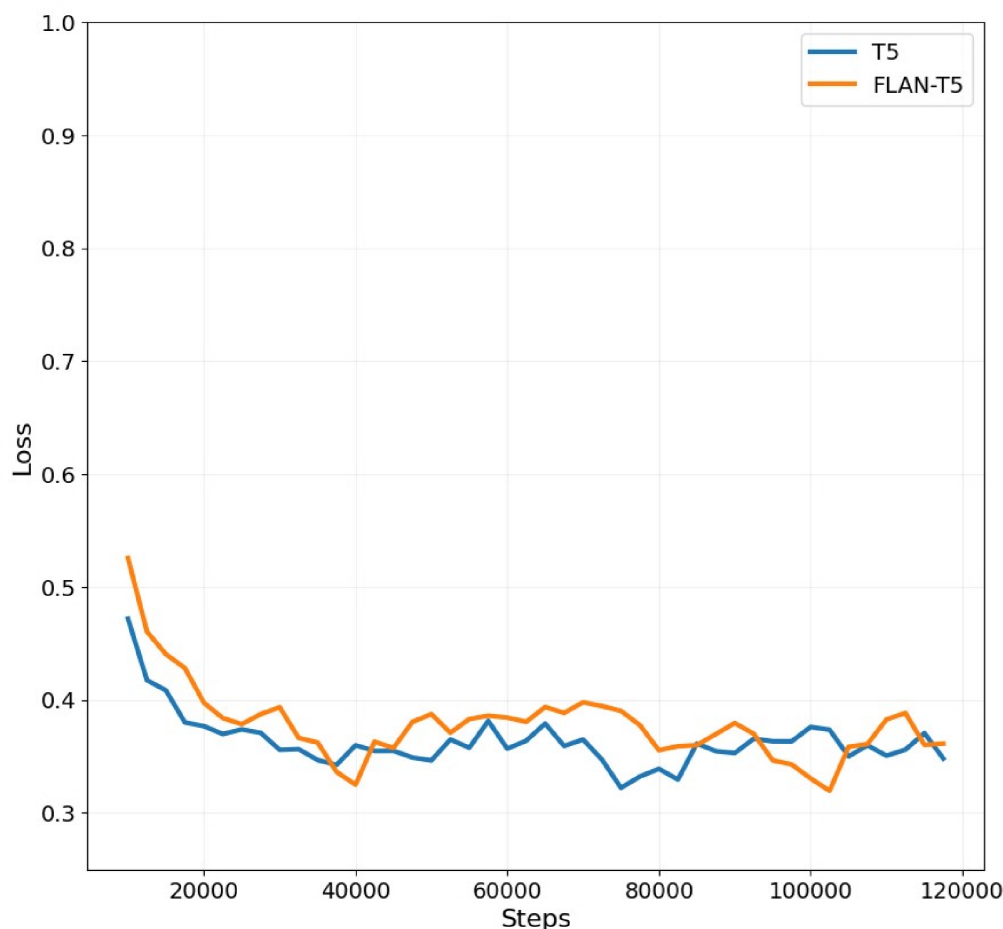


Figure 2. Training loss curves for T5 and Flan-T5 on the GEC task.

Table 4. Overall CEFR Classification Performance

Model	Accuracy (%)	Weighted F1	κ
BERT	97.08	0.96	0.89
RoBERTa	97.13	0.97	0.88
DistilBERT	96.47	0.94	0.84

extremely low-frequency and semantically complex categories. This gap between strong aggregate agreement and weaker tail-class performance highlights the importance of reporting per-level CEFR results in addition to weighted summary metrics. This suggests that while the model effectively captures lexical and syntactic indicators of proficiency, it still underrepresents discourse coherence, rhetorical structure, and stylistic sophistication, which are essential for advanced-level CEFR assessment. Nevertheless, the overall agreement remains high at the aggregate level, suggesting that the model is reliable for most proficiency bands while still requiring improvement for the most advanced class.

5.4. Ensemble-Level integration

The confidence-weighted ensemble integrates the outputs of the GEC, Tense Prediction, and CEFR modules (Equation 19). Table 6 reports the aggregate metrics across all subtasks.

To illustrate the interpretability of the ensemble output, Table 7 presents an example combining grammatical correction, tense validation, and CEFR prediction.

Compared with non-integrated module outputs, the ensemble improved BLEU by 1.2%, F1 by 0.5%, and Cohen's Kappa by 0.02 (all with $p < 0.05$). These results confirm that cross-task interaction yields systemic gains: grammatical normalization improves tense reliability, which in turn stabilizes CEFR grading.

This bottom-up reinforcement aligns with human cognitive acquisition—learners first master syntactic structure (GEC), then temporal coherence (tense), and finally communicative competence (CEFR). Technically, the findings demonstrate that hierarchical multi-task transformers enable emergent linguistic intelligence through shared contextual embeddings.

5.5. Pedagogical and technical implications

The results confirm several important implications:

- **Hierarchical Task Synergy:** Grammatical accuracy supports more reliable tense recognition, which in turn contributes to stronger CEFR prediction, reflecting the staged progression of language development.

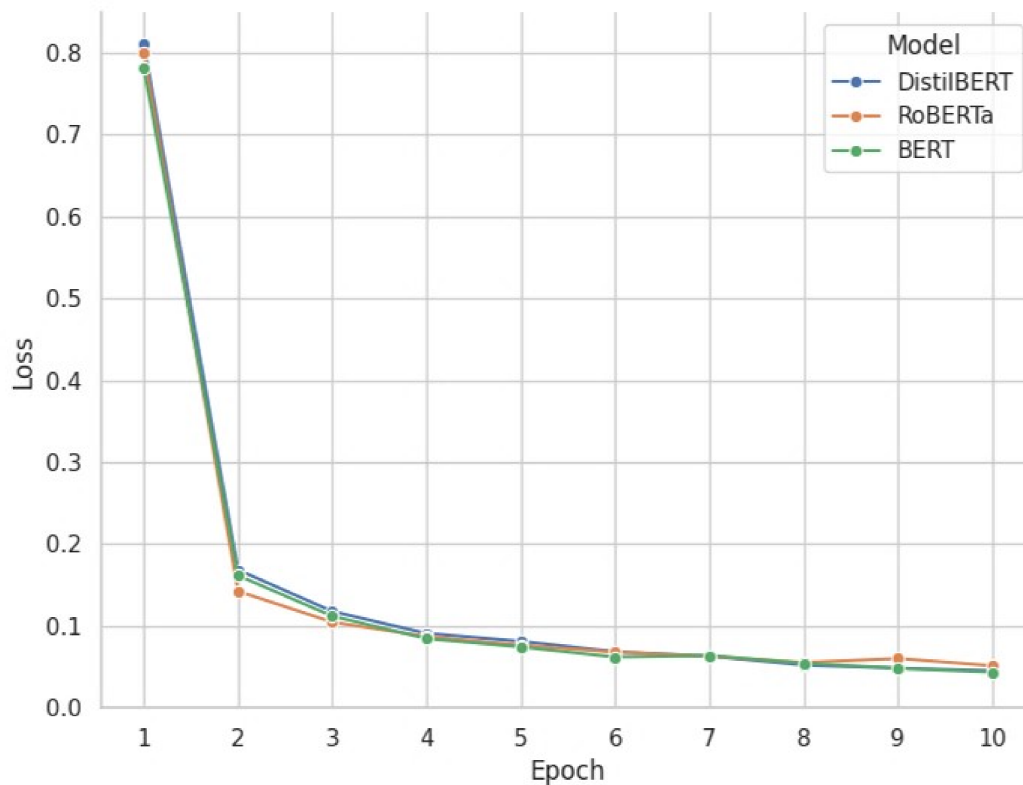


Figure 3. Training loss and F1 trajectories for BERT, RoBERTa, and DistilBERT.

Table 5. Per-Level F1-Score for RoBERTa CEFR Classifier

CEFR Level	F1-score
A1	0.87
A2	0.97
B1	0.98
B2	0.98
C1	0.95
C2	0.57

- **Statistical Robustness:** Most aggregate metrics improved over the compared baseline settings, with statistically significant gains reported for key ensemble-level comparisons ($p < 0.05$), while $\kappa \geq 0.88$ indicates strong agreement with educational reliability standards.
- **Model Efficiency:** Flan-T5 reduces training time by 36.8% while improving BLEU.
- **Pedagogical Utility:** Integrated feedback can be interpretable and actionable for learners.
- **Limitations:** C2-level sensitivity remains challenging due to discourse complexity; future work may incorporate semantic/argumentation modeling.

5.6. Summary

With task-specific peak results of BLEU = 32.4 for grammatical error correction, F1 = 0.99 for tense prediction, and $\kappa = 0.90$ for CEFR agreement, the findings suggest that multi-task transformer architectures can provide linguistically accurate, interpretable, and pedagogically aligned AI feedback.

6. Conclusion and Future Work

This paper presented an AI-powered integrated framework for English language learning that unifies Grammatical Error Correction (GEC), Tense Prediction, and CEFR-Level Proficiency Grading into a single, interpretable transformer-based system. By combining Flan-T5, BERT, RoBERTa, and DistilBERT architectures, the framework demonstrated strong overall linguistic accuracy, interpretability, and computational efficiency within the proposed integrated learning pipeline.

Experimental results validated the effectiveness of the framework. The Flan-T5 module improved grammatical precision (BLEU = 32.4) while reducing training time by 36.8%. The RoBERTa tense model achieved a macro-F1 score of 0.99 with minimal variance ($\sigma = 0.003$). The CEFR classifier reached a Cohen's Kappa of $\kappa = 0.90$, exceeding the 0.80 threshold commonly associated with strong educational rating agreement. The confidence-weighted ensemble further improved BLEU, F1, and κ scores ($p < 0.05$), confirming statistically significant gains and strong synergy among the modules.

From a linguistic standpoint, these findings support a hierarchical learning structure in which grammatical correction improves syntactic clarity, thereby facilitating more accurate tense recognition and, subsequently, more reliable proficiency assessment. This hierarchical alignment between

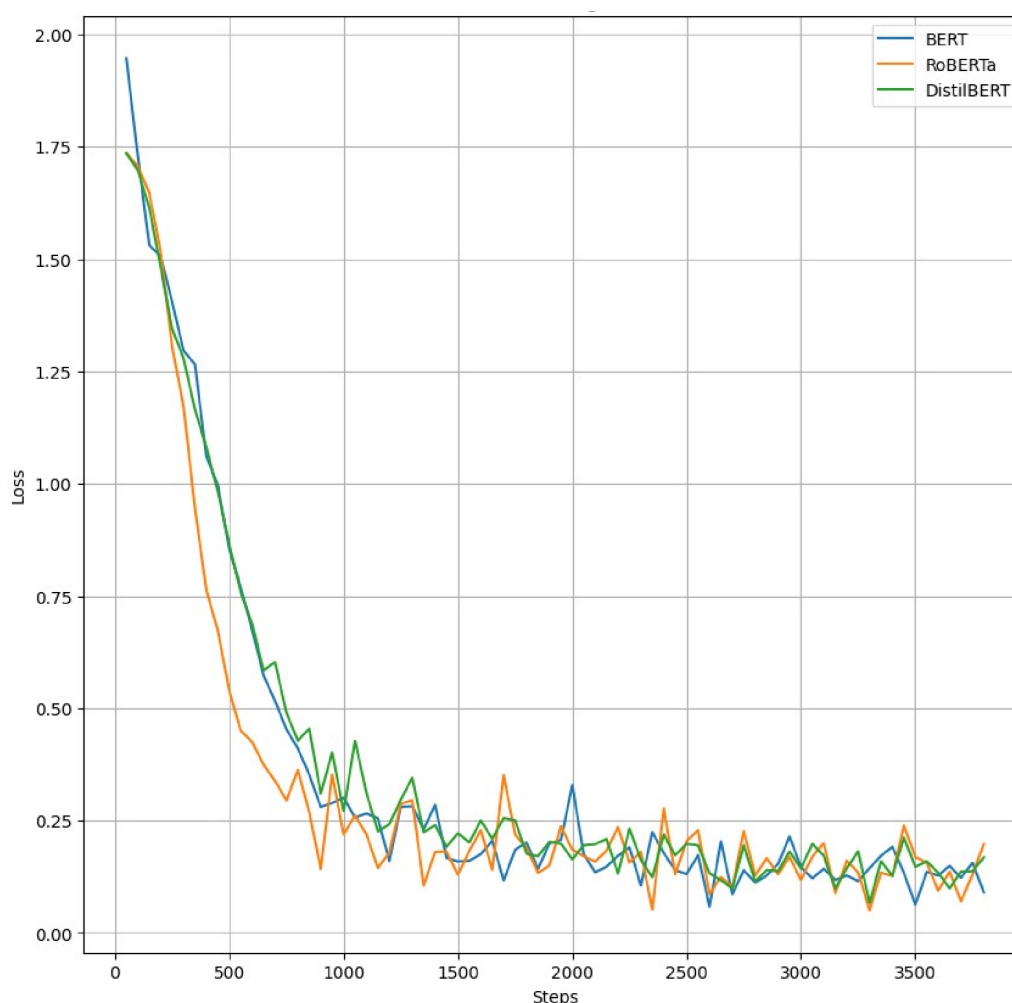


Figure 4. Training loss and confusion matrix for CEFR classification.

Table 6. Integrated Ensemble Results Across All Modules

Metric	Value
BLEU	32.48
F1	0.98
κ	0.90

model design and staged language development positions the framework as a pedagogically interpretable AI assessment system, bridging natural language processing and second-language education.

6.1. Limitations

Although the proposed system achieved strong aggregate results, several limitations should be acknowledged. First, although class-balanced weighting was introduced to mitigate label imbalance, the CEFR classifier remained less reliable for the most advanced proficiency band, particularly C2. This suggests that reweighting alone is insufficient when proficiency distinctions depend heavily on discourse-level and pragmatic competence rather than only sentence-level lexical and syntactic features. As a result, extreme class sparsity and the semantic complexity of advanced learner writing remain challenging for the current modeling framework. Second, the current system relies exclusively on text-based inputs and does not incorporate multimodal signals such as speech, prosody, or interactional context. Third, long-term real-world deployment will require systematic bias mitigation, fairness monitoring, and model drift control to ensure stable and interpretable performance across diverse learner populations.

6.2. Future work

Future research should extend the proposed framework in several important directions. A primary priority is to improve CEFR prediction at advanced proficiency levels, especially the most advanced proficiency level (C2), through discourse-aware and pragmatically informed modeling. Promising directions include incorporating rhetorical structure information, semantic coherence indicators, argumentation-aware representations, and paragraph-level contextual modeling to better capture the organizational and stylistic properties of high-level learner writing. In addition, alternative imbalance-handling strategies, such as focal loss, targeted augmentation for rare advanced-level samples, and curriculum-based fine-tuning, should be explored to improve robustness for extremely low-frequency classes.

Another important direction is to move beyond text-only analysis by integrating multimodal learner signals, such as speech, prosody, fluency,

Table 7. Example of Ensemble Feedback Combining GEC, Tense Validation, and CEFR Prediction

Learner Input	System Feedback
Original Sentence	She go to school every day.
GEC Correction	She goes to school every day.
Tense Validation	Simple Present (consistent)
CEFR Prediction	A2 Level (Basic User)
Integrated Feedback	“Excellent use of present tense. Focus on subject–verb agreement and expand sentence variety to reach B1 level.”

and interactional context. Such extensions would support a more comprehensive assessment of communicative competence and better reflect authentic language use in educational environments.

Future studies should also evaluate the framework through longitudinal and real-world classroom deployment. This includes examining performance stability over time, validating adaptation across diverse learner populations, and developing mechanisms for bias auditing, fairness monitoring, and drift-aware model updating. These efforts are necessary to ensure that the system remains reliable, interpretable, and pedagogically meaningful in practical learning settings.

6.3. Final remarks

In summary, this work contributes to both technical natural language processing and pedagogical practice by demonstrating that multi-task transformer ensembles can model interdependent linguistic skills in an interpretable, scalable, and data-driven manner. The proposed system lays a solid foundation for next-generation intelligent tutoring systems capable of delivering adaptive, explainable, and equitable language learning experiences.

Article Information Form

Author’s Contributions: All authors contributed significantly to the development of this research. *Conceptualization:* Dung Quoc Ho, Lien Thi Quynh Le. *Methodology:* Dung Quoc Ho, Hung Thanh Le. *Software:* Hung Thanh Le, Vu Dang Minh Tran. *Validation:* Dung Quoc Ho, Vu Dang Minh Tran. *Formal Analysis:* Dung Quoc Ho, Lien Thi Quynh Le. *Investigation:* Hung Thanh Le, Vu Dang Minh Tran. *Data Curation:* Vu Dang Minh Tran. *Writing – Original Draft:* Dung Quoc Ho. *Writing – Review & Editing:* Lien Thi Quynh Le. *Visualization:* Hung Thanh Le. *Supervision:* Dung Quoc Ho. *Project Administration:* Lien Thi Quynh Le.

The corresponding author is Lien Thi Quynh Le (lethiquynhlien@hueuni.edu.vn).

Acknowledgments: The authors would like to thank colleagues and students from the Advanced Knowledge Application Laboratory (AKA Lab), Hue University, for their constructive discussions and support during the development of this research.

Conflict of Interest: The authors declare that there is no conflict of interest regarding the research, authorship, or publication of this article.

Support Section: This research did not receive any financial support, grants, or funding from institutions or organizations.

Ethical Approval: This study does not involve human subjects, interviews, surveys, or experiments requiring institutional ethics approval. The authors confirm that all scientific and ethical principles were followed and that all referenced materials are properly cited.

Availability: The datasets used in this study were obtained from publicly available benchmark sources cited in the manuscript. The processed data, code, and experimental configuration files are available from the corresponding author upon reasonable request for research purposes.

Artificial Intelligence Statement: AI tools were used only for improving grammar, clarity, and language quality. All scientific ideas, modeling decisions, analyses, and interpretations were developed solely by the authors.

Plagiarism Statement: This manuscript has been screened for originality using iThenticate™ and meets the journal’s plagiarism standards.

References

- [1] I. Kapounová, “Predicting item difficulty by applying machine learning algorithms using item text features,” Ph.D. dissertation, Charles Univ., Prague, Czech Republic, Sep. 2025. [Online]. Available: <https://dspace.cuni.cz/handle/20.500.11956/203312>
- [2] P. Siripol, S. Rhee, S. Thirakunkovit, and A. Liang-Itsara, “Evaluating the consistency of automated CEFR analyzers: a study of English language text classification,” *Int. J. Eval. Res. Educ. (IJERE)*, vol. 14, no. 4, pp. 3283–3294, Aug. 2025, doi: 10.11591/ijere.v14i4.33528.
- [3] S. M. Marier, X. Chen, L. Zhu, and X. Kong, “Grammatical error correction for low-resource languages: A review of challenges, strategies, computational and future directions,” *PeerJ Comput. Sci.*, vol. 11, p. e3044, Jul. 2025, doi: 10.7717/peerj-cs.3044.
- [4] A. Katinskaia, “Assessing Learner Answers in Computer-Aided Language Learning,” Ph.D. dissertation, Dept. Comput. Sci., Univ. Helsinki, Helsinki, Finland, 2025. [Online]. Available: <http://urn.fi/URN:ISBN:978-952-84-1303-5>
- [5] A. Vaswani et al., “Attention is all you need,” in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008, doi: 10.48550/arXiv.1706.03762.
- [6] A. Koyama, M. Mita, S.-Y. Yoon, Y. Takama, and M. Komachi, “Targeted syntactic evaluation for grammatical error correction,” in *Proc. 63rd Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2025, pp. 21108–21125, doi: 10.18653/v1/2025.acl-long.1026.
- [7] M. Gu, “Cross-lingual pre-trained models for Chinese-English QA systems,” in *Proc. 2nd Int. Conf. Mach. Intell. Digit. Appl. (MIDA)*, 2025, pp. 128–137, doi: 10.1145/3744464.3744485.
- [8] M. Qiu et al., “Chinese grammatical error correction: A survey,” arXiv:2504.00977, Apr. 2025, doi: 10.48550/arXiv.2504.00977.
- [9] X. Zheng and J. Zhang, “The usage of a transformer-based and artificial intelligence-driven multidimensional feedback system in English writing instruction,” *Scientific Reports*, vol. 15, Art. no. 19268, 2025, doi: 10.1038/s41598-025-05026-9.
- [10] C. Lin, “Design and application of an English smart teaching system based on intelligent large language model,” in *Proc. ICICKE 2025*, 2025, pp. 1–6, doi: 10.1109/ICICKE65317.2025.11136776.
- [11] X. Mei, “Visual-textual co-attention model for adaptive English teaching resource recommendation,” in *Proc. ICDSIS 2025*, 2025, pp. 1–9, doi: 10.1109/ICDSIS65355.2025.11070698.

- [12] C. Bryant, M. Felice, Ø. E. Andersen, and T. Briscoe, "The BEA-2019 shared task on grammatical error correction," in *Proc. 14th Workshop Innovative Use of NLP for Building Educational Applications*, 2019, pp. 52–75, doi: 10.18653/v1/W19-4406.
- [13] C. Raffel *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2023, doi: 10.48550/arXiv.1910.10683.
- [14] M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz, "Building a large annotated corpus of English: The Penn Treebank," *Comput. Linguist.*, vol. 19, no. 2, pp. 313–330, 1993.
- [15] *The British National Corpus (BNC)*, Oxford Univ., 1995. [Online]. Available: <https://www.gloriacappelli.it/wp-content/uploads/2007/05/bnc.pdf>
- [16] J. Geertzen, T. Alexopoulou, and A. Korhonen, "Automatic linguistic annotation of large-scale L2 databases: The EF-Cambridge Open Language Database (EFCamDat)," in *Proc. 31st Second Language Research Forum (SLRF)*, Cascadia Proceedings Project, 2013, pp. 240–254.
- [17] *Cambridge Learner Corpus*, Cambridge Univ. Press. [Online]. Available: https://www.cambridge.org/elt/corpus/learner_corpus2.htm
- [18] N. Lin *et al.*, "A BERT-based unsupervised grammatical error correction framework," arXiv:2303.17367, Mar. 2023, doi: 10.48550/arXiv.2303.17367.
- [19] Y. Liu *et al.*, "RoBERTa: A robustly optimized BERT pretraining approach," arXiv:1907.11692, Jul. 2019, doi: 10.48550/arXiv.1907.11692.
- [20] G. Lample and A. Conneau, "Cross-lingual language model pretraining," arXiv:1901.07291, Jan. 2019, doi: 10.48550/arXiv.1901.07291.
- [21] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," arXiv:1711.05101, Jan. 2019, doi: 10.48550/arXiv.1711.05101.
- [22] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2002, pp. 311–318, doi: 10.3115/1073083.1073135.
- [23] C.-Y. Lin, "ROUGE: a package for automatic evaluation of summaries," in *Text Summarization Branches Out*, Jul. 2004, pp. 74–81. [Online]. Available: <https://aclanthology.org/W04-1013/>
- [24] D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," arXiv:2010.16061, Oct. 2020, doi: 10.48550/arXiv.2010.16061.
- [25] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, 1960, doi: 10.1177/001316446002000104.