Zero-Shot Grammar Competency Estimation Using Large Language Model Generated Pseudo Labels

Sourya Dipta Das, Shubham Kumar, Kuldeep Yadav
SHL Labs, India
sourya.das@shl.com, shubham.kumar1@shl.com, kuldeep.yadav@shl.com

Abstract

Grammar competency estimation is essential for assessing linguistic proficiency in both written and spoken language; however, the spoken modality presents additional challenges due to its spontaneous, unstructured, and disfluent nature. Developing accurate grammar scoring models further requires extensive expert annotation, making large-scale data creation impractical. To address these limitations, we propose a zero-shot grammar competency estimation framework that leverages unlabeled data and Large Language Models (LLMs) without relying on manual labels. During training, we employ LLM-generated predictions on unlabeled data by using grammar competency rubric-based prompts. These predictions, treated as pseudo labels, are utilized to train a transformer-based model through a novel training framework designed to handle label noise effectively. We show that the choice of LLM for pseudo-label generation critically affects model performance and that the ratio of clean-to-noisy samples during training strongly influences stability and accuracy. Finally, a qualitative analysis of error intensity and score prediction confirms the robustness and interpretability of our approach. Experimental results demonstrate the efficacy of our approach in estimating grammar competency scores with high accuracy, paving the way for scalable, low-resource grammar assessment systems.

1 Introduction

Grammar competency assessment is a critical component of assessing language proficiency with wide-ranging applications in education, language learning platforms, automated speech scoring systems, and conversational AI (Vajjala and Meurers, 2016; Burstein et al., 2004; Zechner et al., 2009b; Litman and Silliman, 2004). Accurate grammar competency assessment is essential for understanding the linguistic capabilities of individuals across both written and spoken forms of communication (Vajjala and Meurers, 2016; Chapelle and Chapelle, 2001). However, traditional approaches to grammar assessment are often constrained by their reliance on manually annotated datasets and supervised learning paradigms, which demand significant human expertise and resources for dataset creation (Yannakoudakis et al., 2011a; Bryant et al., 2017). These methods also struggle to scale effectively to diverse linguistic contexts and modalities(Zhao et al., 2024). In recent years, advances in machine learning, particularly with the advent of Large Language Models (LLMs) such as GPT, have enabled significant progress in natural language understanding and generation tasks(Brown et al., 2020; Radford et al., 2019; Devlin et al., 2019). LLMs have demonstrated remarkable capabilities in few-shot and zero-shot learning, allowing them to generalize to new tasks with minimal or no labeled data(Radford et al., 2019; Gao et al., 2020). However, leveraging LLMs for grammar competency evaluation remains underexplored, especially in scenarios where labeled datasets are unavailable or infeasible to create.

In this paper, we introduce a novel zero-shot grammar competency score estimation method that addresses the challenges of traditional grammar assessment approaches. Unlike conventional supervised methods, our approach eliminates the dependency on labeled training data by leveraging unlabeled data in conjunction with LLM-generated predictions. Specifically, during training, we use a grammar competency rubric-based prompt created by language experts to guide the LLM in generating predictions for the grammar competency of unlabeled responses. These predictions serve as a form of pseudo-labels, providing the supervisory signal required to train a transformer-based model. To effectively handle the noise in these labels, we propose a novel training framework designed to maximize the learning potential of the model while ensuring robustness and generaliza-

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 3167–3179

tion. Our method is designed to work effectively across both written and spoken responses, making it versatile in addressing the needs of diverse real-world scenarios. For example, it can be applied to assess written essays, transcribed spoken responses, or other forms of language data, thus bridging the gap between text-based and audio-derived inputs. This adaptability makes our approach highly suitable for diverse language assessment tasks. The key contributions of the proposed work are as follows:

- We propose a method that eliminates the reliance on labeled data by leveraging unlabeled data and LLM-generated predictions, offering a scalable and resource-efficient solution for grammar assessment.
- We design grammar competency rubric-based prompts to guide LLMs in generating predictions aligned with human evaluation criteria, ensuring that the pseudo labels reflect meaningful linguistic features.
- We introduce a novel adaptive sample-weighting-based training framework that effectively utilizes pseudo-labels to train a transformer-based model, ensuring robustness and minimizing the impact of label noise.
- Our method supports both written and spoken responses, demonstrating adaptability across different input modalities and real-world scenarios.
- We introduce two in-house industrial datasets, SGAD and WGAD, and conduct comprehensive experiments on them to rigorously validate the effectiveness of our approach, demonstrating its capability to reliably assess grammar competency in zero-shot settings without reliance on labeled training data.

The proposed method represents a notable advancement in automated grammar assessment. Furthermore, the ability to generalize across written and spoken responses makes our approach particularly valuable for applications in education, where multimodal input is common.

2 Related Work

Automated grammar assessment has primarily evolved along two lines: grammatical error detection/correction (GED/GEC) and holistic proficiency scoring (e.g., CEFR-based). However, fine-grained grammar scoring aligned to rubric-based scales, especially for spoken language, remain underexplored. Recent work has begun bridging this gap by leveraging neural and LLM-based models. For instance, (Kopparapu et al., 2024) introduce a grammar scoring system robust to ASR noise, while (Bannò et al., 2024a) employ Whisper-based models for end-to-end GEC, incorporating disfluency. Other studies such as (Caines et al., 2020; Knill et al., 2019) develop spoken GED using sequence labeling, though they report lower accuracy compared to written tasks. Feature-based methods like (Bannò and Matassoni, 2022) predict spoken proficiency from written grammar errors, and (Lu et al., 2020) explore integrating acoustic cues. Broader surveys (Soni and Thakur, 2018; Tetreault and Leacock, 2014) highlight error categorization and real-time challenges. Despite progress, most approaches remain supervised; to the best of our knowledge, there are currently no zero-shot methods specifically designed for rubric-aligned grammar scoring, particularly in the spoken domain, making this an open and impactful research direction.

2.1 Related Work on Grammar Competency Scoring

Recent work in automated grammar scoring for spoken content has explored diverse strategies to handle the variability of learner speech. POS-based similarity measures and syntactic features have proven effective in capturing grammatical proficiency, especially on short utterances (Yoon and Bhat, 2018; Zechner et al., 2017). Multi-task learning with auxiliary tasks like POS-tagging and native language prediction improves model performance on ASR-transcribed speech (Craighead et al., 2020). Systems like SpeechRaterSM combine fluency, ASR, and language use features to align well with human scoring (Zechner et al., 2009b), while rate of speech (ROS) offers a fast, though imperfect, proxy for fluency (de Wet et al., 2007). Cross-corpus studies show models trained on written grammar errors can generalize to spoken inputs (Bannò and Matassoni, 2022; Yuan and Briscoe, 2016). To enhance robustness, recent work explores self-supervised speech models (e.g., wav2vec 2.0), adversarial augmentation, and mixture-of-experts architectures (Bannò et al., 2023; Yoon et al., 2019; Papi et al., 2021). Prompt-aware content features, such as lexical overlap, also help improve relevance and scoring accuracy (Evanini et al., 2013).

2.2 Large Language Models (LLMs) in educational assessment.

Recent work has explored the potential of large language models (LLMs), especially GPT-4(Achiam et al., 2023), for automated essay scoring and feedback generation. GPT-4 has shown consistency with human raters in evaluating discourse coherence (Naismith et al., 2023) and can provide analytic scores aligned with CEFR criteria in zero-shot settings (Bannò et al., 2024b). Perplexity measures from LLMs have been proposed as proxies for linguistic competence (Sánchez et al., 2024). Studies also demonstrate that prompting LLMs with multi-trait criteria leads to reliable analytic assessments for graduate-level writing (Wang et al., 2025) and short L2 essays (Yancey et al., 2023). Multi-trait scoring frameworks like MTS (Lee et al., 2024) and RMTS (Chu et al., 2024) improve trait-specific accuracy using structured prompting and rationale generation. Other work highlights that prompt design can enhance both scoring and feedback generation (Stahl et al., 2024), though fine-tuning remains crucial for short-answer scoring tasks (Chamieh et al., 2024). LLMs have also been applied to spoken grammar evaluation by generating test variations robust to ASR noise (Kopparapu et al., 2024).

3 Proposed Method

The proposed method estimates grammar competency without labeled training data by adopting a zero-shot learning paradigm. Large language model (LLM) predictions serve as pseudo-labels to train a transformer-based model. Pseudo-labels are generated using an LLM prompted with a grammar competency rubric—a strategy shown to enhance zero-shot essay scoring and feedback (Evanini et al., 2013; Wang et al., 2023). To handle pseudo-label noise, we employ a robust framework inspired by prior work on learning from noisy, trait-specific supervision (Zhang et al., 2021; Bengio et al., 2009). This approach generalizes to both written and spoken tasks, eliminating costly human annotations while outperforming strong LLM-only baselines in grammar scoring accuracy.

3.1 Pseudo-Label Generation with LLM

The first step in our method is to generate pseudo-labels for the unlabeled dataset using a Large Language Model (LLM), $f_{\mathcal{LLM}}(.)$. Given an unlabeled dataset $\mathcal{D}_{\text{unlabeled}} = \{x_i\}_{i=1}^{N}$, where x_i represents a sample (written or spoken response), we prompt the LLM with a grammar competency rubric-based prompt P to produce predictions. Mathematically, the pseudo-labels y_i^{pseudo} are defined as:

$$y_i^{\text{pseudo}} = f_{\mathcal{LLM}}(x_i, P)$$

Here, P is carefully designed to align with the grammar competency scoring rubric, ensuring that the LLM predictions are meaningful approximations of grammar scores. These predictions, while inherently noisy, serve as the foundation for training the transformer model.

3.2 Training Methodology

Our proposed training strategy focuses on deriving reliable grammatical proficiency estimates from imperfect, noisy data. We adopt a robust training framework for regression using deep neural networks, designed to mitigate the effects of noisy or low-quality data through dynamic sample weighting (Zhang et al., 2021; Han et al., 2018b; Song et al., 2022). Our approach iteratively re-weights training examples per epoch based on their observed losses, promoting learning from "clean" samples while down-weighting potentially noisy outliers (Jiang et al., 2018; Kumar et al., 2010; Wu et al., 2020). Using the generated pseudo-labels, we construct a training dataset $\mathcal{D}_{\text{train}} = \{(x_i, y_i^{\text{pseudo}})\}_{i=1}^{N}$. The pseudo-labels y_i^{pseudo} are treated as noisy labels, as they may not perfectly align with true grammar competency scores. This introduces a critical challenge in the training process, which our framework addresses by leveraging robust loss functions and regularization techniques to mitigate the impact of label noise (Zhang et al., 2021; Song et al., 2022).

We begin by leveraging a pre-trained transformer encoder, such as BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2019), without architectural modification, and add a projection layer to map its contextual embeddings to scalar proficiency scores for the regression task. Specifically, we instantiate a regression model $f_\theta(\cdot)$, parameterized by θ, wherein the transformer-based architecture serves as the feature extractor, and the projection layer outputs the estimated grammar competency score $\hat{y}_i$ for each input x_i:

$$\hat{y}_i = f_\theta(x_i) \tag{1}$$

Recognizing that not all pseudo-labels assigned to samples are equally reliable, we implement a sample selection mechanism guided by training loss dynamics. The core idea behind this approach (Han et al., 2018b; Jiang et al., 2018; Kumar et al., 2010) is that not all training examples contribute equally to effective model development; treating all supervision uniformly risks overfitting to mislabeled or inconsistent data. To address this, after each epoch, we analyze the training loss associated with individual samples: those consistently exhibiting high loss are flagged as potentially noisy or misaligned with the scoring rubric and are accordingly downweighted, while samples with lower and more stable losses, which are more likely to reflect the true learning signal, are upweighted. This dynamic prioritization enables the model to focus on higher-quality supervision, thereby promoting robustness and mitigating the influence of unreliable labels.

At the beginning of the training process (epoch $t = 0$), all samples are assigned equal importance via uniform weights: $w_i^{(0)} = \frac{1}{N}, \quad \forall i \in \{1, \ldots, N\}$ ensuring $\sum_{i=1}^{N} w_i^{(0)} = 1$. This uniform initialization ensures unbiased exposure to all samples during the initial learning phase. To implement dynamic sample selection in subsequent epochs, we compute the per-sample loss at the end of each epoch. Throughout each training epoch, the model predicts scores $\hat{y}_i = f_\theta^{(t)}(x_i)$ for all samples, and the per-sample loss is computed using the mean squared error (MSE) loss function at epoch t:

$$\ell_i^{(t)} = (f_\theta^{(t)}(x_i) - y_i^{pseudo})^2.$$

During mini-batch training, for each batch $\mathcal{B}_k$ at step k, the training loss is computed as the weighted average of per-sample losses in the batch:

$$L_{\text{batch}}^{(k)} = \frac{1}{|\mathcal{B}_k|} \sum_{i \in \mathcal{B}_k} w_i^{(t)} \cdot \ell_i^{(t)}.$$

This weighted approach ensures that samples deemed more reliable (i.e., with higher weights) exert greater influence on model parameter updates, thereby reducing the impact of noisy or mislabeled data. To adapt sample weights in subsequent epochs, we employ a soft selection strategy. At the end of each epoch, all per-sample losses from the current model parameters are aggregated into a vector:

$$\mathbf{l}^{(t)} = [\ell_1^{(t)}, \ldots, \ell_N^{(t)}].$$

The samples are then sorted in ascending order of their loss values:

$$\pi = argsort(\mathbf{l}^{(t)}).$$

so that samples with the lowest losses, those which the model currently identifies as most "clean," confident, or consistent with the target signal, appear first. Only the top fraction α (e.g., the top 30%) of these samples are retained for the subsequent epoch:

$$I_{\text{clean}}^{(t)} = \{\pi_1, \ldots, \pi_{\lfloor \alpha N \rfloor}\}, \quad 0 < \alpha < 1.$$

These selected samples guide the learning process in the next epoch. Crucially, this dynamic process continuously adjusts sample weights: emphasizing reliable data while still allowing uncertain examples to re-enter training in future epochs as their losses improve. The updated sample weights for epoch $t + 1$ are assigned as follows:

$$w_i^{(t+1)} = \begin{cases} \frac{1}{|I_{\text{clean}}^{(t)}|}, & \text{if } i \in I_{\text{clean}}^{(t)} \\ 0, & \text{otherwise} \end{cases}$$

with normalization to ensure $\sum_{i=1}^{N} w_i^{(t+1)} = 1$. Unlike hard filtering, this dynamic reweighting does not permanently exclude higher-loss samples; instead, it allows their reinclusion in subsequent epochs if their loss improves, capturing the evolving confidence and understanding of the model. The ultimate training objective, given the epoch-wise sample weighting, is to minimize the overall weighted loss:

$$\theta^* = \arg\min_\theta \sum_{i=1}^{N} w_i^{(t)} \ell_i^{(t)}.$$

Here, the adaptive weights $w_i^{(t)}$ dynamically shift focus towards the most informative and consistent samples as determined by model predictions at each stage, facilitating robust training in the presence of label noise and enhancing the overall performance of the grammar score predictor.

4 Experimentation and Results

4.1 Dataset Details

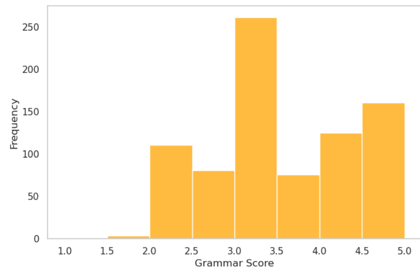
Due to the lack of open-source datasets featuring grammar proficiency ratings, we constructed two in-house datasets to evaluate the performance of our proposed method. These datasets are designed to assess grammar proficiency in both spoken and written modalities, with one dataset for each modality. Each consists of spontaneous speech and written essays, respectively, collected from a diverse participant pool representative of various demographic factors, including gender, region, and linguistic background. Both datasets are divided into two splits: (1) an unlabeled training set, and (2) a test set with ground truth ratings assigned by expert human raters, which we utilize for evaluation metrics. During data collection, we ensured that there was no overlap between participants in the training and test sets, and that the test sets exhibited no significant class imbalance. The distribution of ratings for each dataset is presented in Table 1. Further details on each dataset are provided below.

Spoken Grammar Assessment Dataset (SGAD) : The SGAD dataset was derived from an online spoken English assessment product, where candidates responded spontaneously to two open-ended prompts, each within a 60-second time limit. Prompts were designed to elicit natural language use and authentic grammatical structures. All audio responses were transcribed using a state-of-the-art automatic speech recognition (ASR) system[1] for accurate textual representation. . For the test set, four expert raters, representing diverse linguistic backgrounds and possessing expertise in language assessment, evaluated both audio and transcripts with Subject Matter Experts (SME) . This rubric assessed grammatical accuracy, fluency, and coherence. Each response was rated by multiple experts to ensure reliability, with final scores averaged to address inter-rater variability.

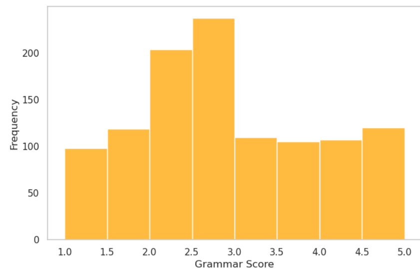
Written Grammar Assessment Dataset (WGAD) : The WGAD was developed using an analogous methodology, leveraging an online language assessment product intended to evaluate written English proficiency. In this test, participants were required to write structured essays on given topics, facilitating the assessment of grammar use in formal writing contexts. For the test set, essays were evaluated by expert raters using a specialized five-point rubric for written grammar, also devised by I/O psychologists and linguists to assess grammatical accuracy, coherence, and fluency. The rater panel consisted of four individuals with diverse demographic and linguistic backgrounds to ensure robust and unbiased evaluation. Multiple experts rated each essay, and discrepancies were resolved through score averaging. Inter-rater correlation was computed to validate the reliability of the ratings.

[1]We used the Azure Speech to Text service by Microsoft (https://learn.microsoft.com/en-us/azure/ai-services/speech-service/speech-to-text) for transcribing all audio data.

Unlabeled Training Data Preparation and Pseudo-Labeling for SGAD and WGAD) : To construct the unlabeled training datasets for both SGAD and WGAD, we collected extensive spoken and written samples, respectively, from over 10,000 individuals representing a broad spectrum of linguistic and regional backgrounds, thereby mitigating potential demographic bias during training. Each participant provided two responses to assigned prompts or topics, resulting in large, demographically diverse corpora for both modalities. For both SGAD and WGAD, pseudo-labels were generated using the GPT-4 (OpenAI, 2023) model, which was prompted with the same five-point grammar scoring rubrics employed by human raters, specifically, the spoken grammar rubric for SGAD and the written grammar rubric for WGAD, to assign scores ranging from 1 to 5. This approach ensured consistency with human evaluation standards, reduced subjectivity, and addressed the scalability limitations inherent in manual annotation. For each dataset, additional details regarding the train and test splits are provided in Table 1.



(a)



(b)

Figure 1: Histogram Plot of Expert-Rated Grammar Scores from Test Set of (a) SGAD Dataset (b) WGAD Dataset

4.2 Evaluation Metric

To rigorously assess the performance of the grammar competency scoring model, we employ several evaluation metrics commonly used in related research (Yannakoudakis et al., 2011b; Zechner et al., 2009a; Williamson et al., 2012; Attali and Burstein, 2006). Specifically, we report the Quadratic

Weighted Kappa (QWK), the Pearson Linear Correlation Coefficient (PLCC), the Spearman Rank Correlation Coefficient (SRCC), and the Root Mean Square Error (RMSE). QWK evaluates agreement between predicted and expert-annotated scores, and making it well-suited for ordinal regression. PLCC measures the linear correlation between predicted and reference scores, while SRCC assesses the consistency of rank ordering, capturing both linear and non-linear monotonic relationships. RMSE quantifies prediction error as the square root of the mean squared differences between predicted and actual scores. During evaluation, higher values of QWK, PLCC, and SRCC, alongside a lower RMSE, are indicative of better grammar competency scoring performance.

Table 1: Details of the Grammar Assessment Datasets.

Dataset	Dataset Split	No of Samples	No of Candidates	Avg. Length (Words)	Max Length (Words)
SGAD	Train-set	20030	10015	74	130
	Test-set	778	389	72	132
WGAD	Train-set	9669	9669	260	539
	Test-set	1059	1059	258	422

4.3 Performance of Different Backbone Model Architectures

We evaluated three backbone architectures: BERT, ELECTRA, and XLNet on the SGAD and WGAD datasets and results are shown in Table 2. ELECTRA consistently outperformed the others, achieving the highest QWK, PLCC, and SRCC scores with the lowest RMSE across both datasets. On SGAD, it showed the strongest agreement with human ratings, while also maintaining robust correlation scores. On WGAD, ELECTRA continued to lead, confirming its effectiveness across modalities. BERT followed closely, particularly in WGAD, with competitive QWK and PLCC scores, though its higher RMSE and slightly lower correlations suggest minor prediction inconsistencies. XLNet trailed both models, with lower agreement metrics and higher RMSE, indicating limited suitability for grammar scoring tasks. Overall, ELECTRA's performance highlights the value of its pretraining approach and underscores the importance of selecting strong transformer models for reliable grammar assessment.

4.4 Performance of Different LLM Model Architectures

We evaluate several large language models (LLMs) on the SGAD and WGAD datasets to measure their ability to predict grammar proficiency (Table 3). For score prediction, we apply the same grammatical competency rubric-based prompt that was used during pseudo-label generation. The models differ in architecture and training methods, providing insights into what works best for spoken and written grammar assessment. Most models perform well on written grammar, showing strong correlation with expert scores. However, performance drops in spoken grammar tasks, where disfluencies and spontaneous speech are harder to handle. Models trained with task-specific data tend to perform more reliably. Among all models, GPT-4 consistently outperforms other LLMs across both spoken and written grammar evaluations, further establishing it as a baseline and a suitable choice for use in the proposed method. Overall, results highlight the need for careful model selection and targeted training for grammar evaluation.

4.5 Method Sensitivity to Different LLM Model Architectures

We conducted a comprehensive study using the top five performing large language models (LLMs) listed in Table 3 to evaluate the sensitivity of our approach to variations in LLM architectures. Each model generated pseudo labels for grammar scoring under identical instructions and evaluation rubrics to ensure a controlled comparison. We then trained separate instances of our grammar scoring model on the pseudo-labeled datasets from each LLM, employing the optimal configuration identified in previous experiments. Results are reported in Table 4. Our analysis reveals that downstream model performance is strongly influenced by the capability of the LLM used to produce pseudo labels. Models trained on labels from higher-capability LLMs, those demonstrating stronger alignment with human-rated grammar scores, exhibited superior agreement with expert annotations. Conversely, pseudo labels generated by less capable LLMs introduced higher noise, leading to reduced performance. This dependency reflects the ability of advanced LLMs to capture nuanced grammatical features and produce pseudo labels that closely mirror expert judgments. Consequently, the reliability and quality of supervision scale with the underlying LLM's intrinsic proficiency.

4.6 Sensitivity Analysis of the α Parameter

The hyperparameter α plays a critical role in controlling the noise filtering mechanism by determin-

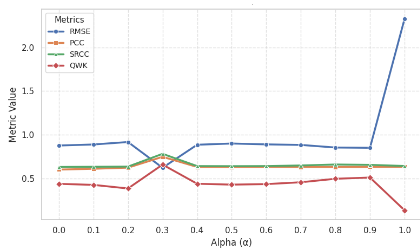
Table 2: Performance of Different Backbone Model Architectures.

Dataset	Model	QWK	PLCC	SRCC	RMSE
SGAD	BERT (Devlin et al., 2019)	0.659	0.748	0.782	0.623
	ELECTRA (Clark et al., 2020)	0.664	0.732	0.73	0.73
	XL-Net (Yang et al., 2019)	0.589	0.623	0.664	0.844
WGAD	BERT (Devlin et al., 2019)	0.776	0.862	0.813	0.558
	ELECTRA (Clark et al., 2020)	0.763	0.833	0.797	0.599
	XL-Net (Yang et al., 2019)	0.664	0.686	0.690	0.912

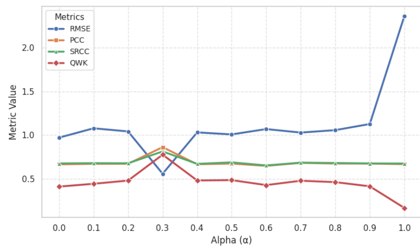
Table 3: Performance of Different LLM Model Architectures on both the WGAD and SGAD datasets.

Method	SGAD				WGAD			
	QWK	PLCC	SRCC	RMSE	QWK	PLCC	SRCC	RMSE
GPT-4 (Achiam et al., 2023)	0.543	0.602	0.621	0.998	0.541	0.632	0.645	1.233
GPT-4o (Hurst et al., 2024)	0.533	0.587	0.592	1.082	0.534	0.654	0.666	1.298
Gemini 1.5 (Team et al., 2024)	0.324	0.422	0.447	0.952	0.261	0.458	0.465	1.195
LLAMA 3 (Grattafiori et al., 2024)	0.445	0.542	0.644	1.312	0.411	0.544	0.586	1.403
Mistral-7b (Jiang et al., 2023)	0.261	0.324	0.375	1.702	0.256	0.318	0.465	1.234
Mistral-8x7b (Jiang et al., 2024)	0.282	0.265	0.345	1.611	0.299	0.478	0.592	1.066
Mistral-large (Jiang et al., 2023)	0.455	0.567	0.687	1.044	0.477	0.576	0.496	1.064
Claude Sonnet (Claude)	0.478	0.553	0.632	1.266	0.495	0.599	0.598	1.052
Claude Haiku (Claude)	0.461	0.592	0.622	1.193	0.498	0.576	0.582	0.989

ing the fraction of samples retained as "clean" after each training epoch. To rigorously evaluate the impact of α on model performance and address concerns regarding its selection, we conducted an extensive sensitivity analysis over α values ranging from 0.0 to 1.0 in increments of 0.1. For each value, we classified the lowest-loss α fraction of samples as clean and assigned them higher sampling weights during the subsequent training epoch, while down-weighting the remaining $(1-\alpha)$ fraction. This approach allows us to systematically explore the trade-off between discarding noisy samples and preserving valuable training data. When $\alpha = 0$, all samples are considered noisy and effectively discarded, resulting in minimal data utilization; conversely, $\alpha = 1$ corresponds to using the entire dataset without any noise filtering. Intermediate values of α enable flexible balancing between noise robustness and data retention. The analysis, illustrated in Fig. 2a and Fig. 2b, reveals that model performance exhibits a clear dependence on α. Notably, moderate values of α (e.g., around 0.3) consistently yield lower root mean squared error (RMSE) and improved correlation metrics across multiple datasets, indicating an optimal balance which empirically justifies our original choice of $\alpha = 0.3$.



(a) Effect of α variation (SGAD Dataset)



(b) Effect of α variation (WGAD Dataset)

Figure 2: Sensitivity of Grammar Scoring model performance to α on both the WGAD and SGAD datasets.

4.7 Quantitative Comparison

For baseline comparison, we introduce two baseline approaches: supervised baseline and unsupervised baseline. For unsupervised baseline, we employ GPT-4 large language model (LLM) for grammar scoring, leveraging its strong zero-shot performance in rubric-aligned assessment tasks. During inference, we used the same grammar competency rubric-based prompt as was utilized during pseudo-label generation, thereby ensuring consistency in prediction criteria. For the supervised baseline, we trained a BERT-based grammar scoring model specifically using the same pseudo-labeled dataset as our proposed method. This baseline was optimized using mean squared error (MSE) loss with identical training configurations, except that no label noise-aware sample weighting was applied. The model was evaluated on the same test set as our proposed approach. Including this supervised baseline provides a more comprehensive context for interpreting the performance gains achieved by our pseudo-label–based training framework. While our rated dataset does not contain enough annotated samples to support a conventional fully supervised baseline with an independent train–test split,

Table 4: Performance Comparison of Grammar Scoring Models trained on Pseudo labels from Different LLM architectures on both the WGAD and SGAD datasets.

Method	SGAD				WGAD			
	QWK	PLCC	SRCC	RMSE	QWK	PLCC	SRCC	RMSE
GPT-4o (Hurst et al., 2024)	0.426	0.643	0.660	0.920	0.419	0.687	0.692	1.140
LLAMA 3 (Grattafiori et al., 2024)	0.365	0.605	0.634	0.897	0.409	0.678	0.697	1.114
Mistral-large (Jiang et al., 2023)	0.489	0.684	0.707	0.778	0.309	0.641	0.657	1.028
Claude Sonnet (Claude)	0.374	0.598	0.605	0.862	0.391	0.669	0.676	0.951
Claude Haiku (Claude)	0.341	0.629	0.634	0.842	0.401	0.639	0.645	0.958

this setup serves as an ablation study, quantifying the benefits of our sample weighting and noise-aware training procedures central to the proposed approach. The performance of the respective baselines is reported in Table 3. The results show that our proposed method outperforms both baselines by substantial margins.

Given the lack of prior work in zero-shot grammar scoring, particularly in the spoken domain, we adopt noise-robust learning algorithms as a principled alternative to supervised methods for handling pseudo-labeled data. To benchmark the effectiveness of our approach, we compare it against several state-of-the-art (SOTA) noise-robust training techniques, including co-teaching (Han et al., 2018a), pseudo-label refinement (Wang et al., 2022), and sample reweighting methods (Feng et al., 2024; Li et al., 2022), evaluated on both the SGAD and WGAD datasets (Table 5). While these methods are designed to mitigate the effects of label noise, they often exhibit limited generalization and inconsistent performance across metrics. In contrast, our structured training framework demonstrates robust and stable results, showing greater resilience to noisy supervision.

4.8 Impact Analysis with Different Error Types

Although grammar scoring models effectively assess grammatical proficiency, their reliability in spoken language remains challenged by informal structures, disfluencies, and pauses (Ting et al., 2010). Without distinguishing acceptable spoken variations from true errors, models may misjudge natural speech or miss actual mistakes, reducing alignment with human evaluations. To analyze this, we construct a synthetic dataset by selecting high-scoring (≥ 4.5) samples from SGAD and WGAD and introducing controlled grammatical errors. Domain-specific errors, such as *spelling*, *verb form*, *tense*, *subject–verb agreement*, *pronouns*, *punctuation*, *prepositions*, *word order*, and *filler words*, are applied following prior work (Wang et al., 2021; Ting et al., 2010). Details of each error type appear below.

- **Filler Word Error**: Use of unnecessary words like "um," "like," or "you know."
- **Redundant Phrases**: Repetition of ideas that makes the sentence wordy.
- **Word Order Error**: Incorrect sequence of words affecting clarity and grammar.
- **Verb Error**: Incorrect verb form disrupting sentence structure.
- **Preposition Error**: Wrong or missing prepositions leading to awkward expressions.
- **Tense Errors**: Inconsistent or incorrect verb tenses confusing the time of action.
- **Subject-Verb Agreement Error**: Mismatch in number between subject and verb.
- **Spelling Error**: Incorrect spelling affecting readability or meaning.
- **Punctuation Error**: Misuse or omission of punctuation changing sentence meaning.
- **Pronoun Error**: Unclear use of pronouns confusing the sentence subject or object.

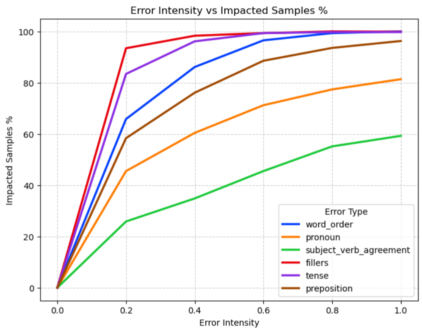
Each sample is corrupted in a controlled manner, where we incrementally increase the *error intensity*, defined as the percentage of words affected. This enables fine-grained stress testing of model robustness across varying degrees of linguistic degradation. The resulting dataset facilitates evaluation of model sensitivity, consistency with expert ratings, and bias in error attribution, offering insights into how different error types influence prediction behavior and helping guide the development of more resilient grammar assessment models.

4.9 Qualitative Comparison

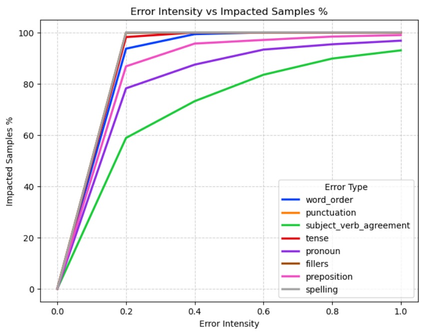
We qualitatively evaluated model robustness using a synthetic dataset (4.8) with varying grammatical error intensities. As errors increased, predicted grammar scores declined, showing a strong negative correlation (Figure 4). The percentage of impacted samples those with higher score differences also increased with error intensity, as shown in Figure 3. Structural errors like word order, filler,

Table 5: Quantitative Comparison Results on both the WGAD and SGAD datasets.

Method	SGAD				WGAD			
	QWK	PLCC	SRCC	RMSE	QWK	PLCC	SRCC	RMSE
Our method	**0.659**	**0.748**	**0.782**	**0.623**	**0.776**	**0.862**	**0.813**	**0.558**
Supervised Baseline	0.466	0.655	0.657	0.997	0.366	0.478	0.488	1.102
Unsupervised Baseline	0.543	0.602	0.621	0.998	0.541	0.632	0.645	1.233
Mentor-net (Jiang et al., 2018)	0.249	0.176	0.141	1.167	0.671	0.821	0.795	0.673
Co-teaching (Han et al., 2018a)	0.155	0.167	0.167	1.386	0.772	0.800	0.795	0.669
Co-teaching Plus (Yu et al., 2019)	0.225	0.412	0.410	2.137	0.766	0.795	0.782	0.677
SIGUA (Han et al., 2020)	0.585	0.733	0.761	0.737	0.580	0.814	0.786	0.667
FINE (Kim et al., 2021)	0.499	0.265	0.253	1.298	0.731	0.817	0.798	0.632
Active-Passive-Losses (Ma et al., 2020)	0.640	0.699	0.738	0.695	0.733	0.802	0.775	0.681
SPR-LNL (Wang et al., 2022)	0.358	0.651	0.667	1.057	0.447	0.533	0.596	1.125
SSR-BMV (Feng et al., 2024)	0.624	0.731	0.742	0.655	0.696	0.803	0.769	0.662
Sel-CL (Li et al., 2022)	0.587	0.712	0.745	0.724	0.756	0.804	0.782	0.658

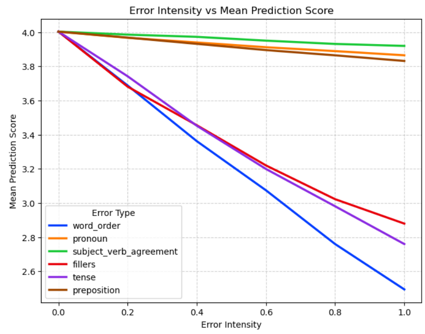


(a)

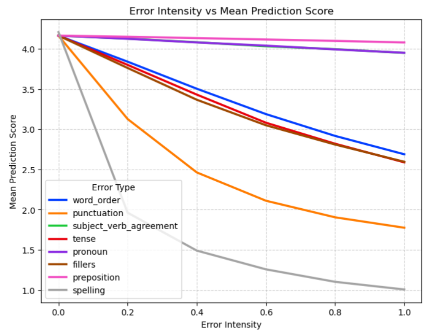


(b)

Figure 3: Percentage of impacted samples due to increasing Error Intensity on Mean Prediction Scores Across Error Types. (a) Impact on SGAD Dataset. (b) Impact on WGAD Dataset



(a)



(b)

Figure 4: Impact of Increasing Error Intensity on Mean Prediction Scores Across Error Types. (a) Impact on SGAD Dataset. (b) Impact on WGAD Dataset

and punctuation caused the largest drops. Comparison with human ratings showed strong alignment, confirming consistent rubric-based scoring.

5 Conclusion

We present a novel zero-shot method to estimate grammar competency scores in both written and spoken responses. Our approach mitigates the scarcity of labeled data by leveraging unlabeled samples and generating pseudo-labels using Large Language Model (LLM) predictions guided by a rubric-based prompt. These pseudo-labels are then employed within a noise-aware training framework to train a transformer-based model for grammar score prediction. The method's ability to generalize across written and spoken modalities demonstrates its broad applicability. Experimental results highlight the effectiveness of our approach and its ability to overcome the limitations of labeled data scarcity. Additionally, experiments varying the ratio of "clean" and "noisy" samples retained after each epoch reveal that selective retention of high-quality samples is crucial for stable training. We further evaluate multiple LLMs for pseudo-label generation, showing that model choice significantly influences alignment with human judgment. Future work will enhance noise robustness and extend to multilingual datasets.

Limitations

While our method offers a practical and scalable solution for grammar competency score estimation, it has certain limitations. First, the use of pseudo-labels derived from Large Language Model (LLM) predictions introduces noise and some errors may not be captured well by LLMs, which may affect the model's accuracy under certain conditions. Second, the approach relies on the quality and alignment of the grammar competency rubric-based prompts, which may also vary across different use cases.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater® v.2. *Journal of Technology, Learning, and Assessment*, 4(3):1–21.

Stefano Bannò, Katherine M Knill, Marco Matassoni, Vyas Raina, and Mark Gales. 2023. Assessment of l2 oral proficiency using self-supervised speech representation learning. ISCA.

Stefano Bannò, Rao Ma, Mengjie Qian, Kate M Knill, and Mark JF Gales. 2024a. Towards end-to-end spoken grammatical error correction. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10791–10795. IEEE.

Stefano Bannò and Marco Matassoni. 2022. Cross-corpora experiments of automatic proficiency assessment and error detection for spoken english. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 82–91.

Stefano Bannò, Hari Krishna Vydana, Kate M Knill, and Mark JF Gales. 2024b. Can gpt-4 do l2 analytic assessment? *arXiv preprint arXiv:2404.18557*.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Tom B. Brown, Benjamin Mann, Nick Ryder, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

CJ Bryant, Mariano Felice, and Edward Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. Association for Computational Linguistics.

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

Andrew Caines, Christian Bentz, Kate Knill, Marek Rei, and Paula Buttery. 2020. Grammatical error detection in transcriptions of spoken english. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2144–2162.

Imran Chamieh, Torsten Zesch, and Klaus Giebermann. 2024. Llms in short answer scoring: Limitations and promise of zero-shot and few-shot approaches. In *Proceedings of the 19th workshop on innovative use of nlp for building educational applications (bea 2024)*, pages 309–315.

Carol Chapelle and Carol A Chapelle. 2001. *Computer applications in second language acquisition*. Cambridge university press.

SeongYeub Chu, JongWoo Kim, Bryan Wong, and MunYong Yi. 2024. Rationale behind essay scores: Enhancing s-llm's multi-trait essay scoring with rationale generated by llms. *arXiv preprint arXiv:2410.14202*.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Claude. The claude 3 model family: Opus, sonnet, haiku.

Hannah Craighead, Andrew Caines, Paula Buttery, and Helen Yannakoudakis. 2020. Investigating the effect of auxiliary objectives for the automated grading of learner english speech transcriptions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2258–2269.

Febe de Wet, Christa van der Walt, and Thomas Niesler. 2007. Automatic large-scale oral language proficiency assessment. In *Interspeech*, pages 218–221.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Keelan Evanini, Shasha Xie, and Klaus Zechner. 2013. Prompt-based content scoring for automated spoken language assessment. In *Proceedings of the eighth workshop on innovative use of NLP for building educational applications*, pages 157–162.

Chen Feng, Georgios Tzimiropoulos, and Ioannis Patras. 2024. Noisebox: Towards more efficient and effective learning with noisy labels. *IEEE Transactions on Circuits and Systems for Video Technology*.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Bo Han, Gang Niu, Xingrui Yu, Quanming Yao, Miao Xu, Ivor Tsang, and Masashi Sugiyama. 2020. Sigua: Forgetting may make learning with noisy labels more robust. In *International Conference on Machine Learning*, pages 4006–4016. PMLR.

Bo Han, Jiangchao Yao, Gang Niu, Mingyuan Zhou, Ivor Tsang, Ya Zhang, and Masashi Sugiyama. 2018a. Masking: A new perspective of noisy supervision. *Advances in neural information processing systems*, 31.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W. Tsang, and Masashi Sugiyama. 2018b. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 8527–8537.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. 2018. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*.

Taehyeon Kim, Jongwoo Ko, JinHwan Choi, Se-Young Yun, et al. 2021. Fine samples for learning with noisy labels. *Advances in Neural Information Processing Systems*, 34:24137–24149.

Kate M Knill, Mark JF Gales, PP Manakul, and AP Caines. 2019. Automatic grammatical error detection of non-native spoken learner english. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 8127–8131. IEEE.

Sunil Kumar Kopparapu, Chitralekha Bhat, and Ashish Panda. 2024. Spoken grammar assessment using llm. *arXiv preprint arXiv:2410.01579*.

M. Pawan Kumar, Benjamin Packer, and Daphne Koller. 2010. Self-paced learning for latent variable models. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1189–1197.

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

Shikun Li, Xiaobo Xia, Shiming Ge, and Tongliang Liu. 2022. Selective-supervised contrastive learning with noisy labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 316–325.

Diane Litman and Scott Silliman. 2004. Itspoke: An intelligent tutoring spoken dialogue system. In *Demonstration papers at HLT-NAACL 2004*, pages 5–8.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yiting Lu, Mark JF Gales, and Yu Wang. 2020. Spoken language'grammatical error correction'. ISCA.

Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah Erfani, and James Bailey. 2020. Normalized loss functions for deep learning with noisy labels. In *International conference on machine learning*, pages 6543–6553. PMLR.

Ben Naismith, Phoebe Mulcaire, and Jill Burstein. 2023. Automated evaluation of written discourse coherence using gpt-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 394–403.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Sara Papi, Edmondo Trentin, Roberto Gretter, Marco Matassoni, and Daniele Falavigna. 2021. Mixtures of deep neural experts for automated speech scoring. *arXiv preprint arXiv:2106.12475*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Ricardo Muñoz Sánchez, Simon Dobnik, and Elena Volodina. 2024. Harnessing gpt to study second language learner essays: Can we use perplexity to determine linguistic competence? In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 414–427.

H. Song, M. Kim, D. Park, Y. Shin, J. Y. Lee, and J. Lee. 2022. Learning from noisy labels with small-loss selection: A survey. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, volume 45, pages 427–447.

Madhvi Soni and Jitendra Singh Thakur. 2018. A systematic review of automated grammar checking in english language. *arXiv preprint arXiv:1804.00540*.

Maja Stahl, Leon Biermann, Andreas Nehring, and Henning Wachsmuth. 2024. Exploring llm prompting strategies for joint essay scoring and feedback generation. *arXiv preprint arXiv:2404.15845*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Joel Tetreault and Claudia Leacock. 2014. Automated grammatical error correction for language learners. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Tutorial Abstracts*, pages 8–10.

Su-Hie Ting, Mahanita Mahadhir, and Chang Siew-Lee. 2010. Grammatical errors in spoken english of university students in oral communication course. *GEMA Online Journal of Language Studies*, 10(1):53.

Sowmya Vajjala and Detmar Meurers. 2016. Readability-based sentence ranking for evaluating text simplification. *arXiv preprint arXiv:1603.06009*.

Xiaodong Wang, Xiaole Zhou, Matthew Fulton, and Sowmya Vajjala. 2023. Prompting for trait-wise supervision: Using rubrics and llms for automated essay scoring. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 122–134.

Yikai Wang, Xinwei Sun, and Yanwei Fu. 2022. Scalable penalized regression for noise detection in learning with noisy labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 346–355.

Yu Wang, Yuelin Wang, Kai Dang, Jie Liu, and Zhuo Liu. 2021. A comprehensive survey of grammatical error correction. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 12(5):1–51.

Zhengxiang Wang, Veronika Makarova, Zhi Li, Jordan Kodner, and Owen Rambow. 2025. Llms can perform multi-dimensional analytic writing assessments: A case study of l2 graduate-level academic english writing. *arXiv preprint arXiv:2502.11368*.

David M. Williamson, Xiaoming Xi, and Frederick J. Breyer. 2012. Best practices for evaluating automated scoring. In *Automated Scoring of Complex Tasks in Computer-Based Testing*, pages 245–282. Routledge.

Hao Wu, Sungjin Lee, and Ser-Nam Lim. 2020. Top-k training of gans: Improving gan performance by selecting top-k samples. In *International Conference on Learning Representations (ICLR)*.

Kevin P Yancey, Geoffrey Laflair, Anthony Verardi, and Jill Burstein. 2023. Rating short l2 essays on the cefr scale with gpt-4. In *Proceedings of the 18th workshop on innovative use of NLP for building educational applications (BEA 2023)*, pages 576–584.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011a. A new dataset and method for automatically grading esol texts. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 180–189.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011b. A new dataset and method for automatically grading esol texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, pages 180–189.

Su-Youn Yoon and Suma Bhat. 2018. A comparison of grammatical proficiency measures in the automated assessment of spontaneous speech. *Speech Communication*, 99:221–230.

Su-Youn Yoon, Chong Min Lee, Klaus Zechner, and Keelan Evanini. 2019. Development of robust automated scoring models using adversarial input for oral proficiency assessment. In *INTERSPEECH*, pages 1871–1875.

Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. 2019. How does disagreement help generalization against label corruption? In *International conference on machine learning*, pages 7164–7173. PMLR.

Zheng Yuan and Ted Briscoe. 2016. Grammatical error correction using neural machine translation. In *Proceedings of the 2016 conference of the north American Chapter of the Association for computational linguistics: Human language technologies*, pages 380–386.

Klaus Zechner, Derrick Higgins, Xiaoming Xi, and David Williamson. 2009a. Automatic scoring of non-native spontaneous speech in tests of spoken english. In *Speech Communication*, volume 51, pages 883–895.

Klaus Zechner, Derrick Higgins, Xiaoming Xi, and David M Williamson. 2009b. Automatic scoring of non-native spontaneous speech in tests of spoken english. *Speech communication*, 51(10):883–895.

Klaus Zechner, Su-Youn Yoon, Suma Bhat, and Chee Wee Leong. 2017. Comparative evaluation of automated scoring of syntactic competence of non-native speakers. *Computers in Human Behavior*, 76:672–682.

Da Zhang, Xinyang Li, Simon Kornblith, Chen-Yu Wang, et al. 2021. Learning from noisy labels with deep neural networks: A survey. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6606–6615.

Chuanjun Zhao, Meiling Wu, Xinyi Yang, Wenyue Zhang, Shaoxia Zhang, Suge Wang, and Deyu Li. 2024. A systematic review of cross-lingual sentiment analysis: tasks, strategies, and prospects. *ACM Computing Surveys*, 56(7):1–37.