# Overview of CCL24-Eval Task 7: Chinese Essay Fluency Evaluation (CEFE) Task

**Xinlin Zhuang**[1], **Xinshu Shen**[1], **Hongyi Wu**[1], **Man Lan**[1,2*], **Xiaopeng Bai**[2,3], **Yuanbin Wu**[1,2], **Aimin Zhou**[1,2], and **Shaoguang Mao**[4]
[1]School of Computer Science and Technology, East China Normal University
[2]Shanghai Institute of AI for Education, East China Normal University
[3]Department of Chinese Language and Literature, East China Normal University
[4]Microsoft Research Asia
{xinlinzhuang,xinshushen,hongyiwu}@stu.ecnu.edu.cn
{mlan,ybwu,amzhou}@cs.ecnu.edu.cn, xpbai@zhwx.ecnu.edu.cn
{shaoguang.mao}@microsoft.com

## Abstract

This paper presents a detailed review of Task 7 in the CCL24-Eval: the second **C**hinese **E**ssay **F**luency **E**valuation (CEFE). The task aims to identify fine-grained grammatical errors that impair readability and coherence in essays authored by Chinese primary and secondary school students, evaluate the essays' fluency levels, and recommend corrections to improve their written fluency. The evaluation comprises three tracks: (1) Coarse-grained and fine-grained error identification; (2) Error sentence rewriting; and (3) Essay Fluency Level Recognition. We garnered 29 completed registrations, resulting in 180 submissions from 10 dedicated teams. The paper discusses the submissions and analyzes the results from all participating teams.

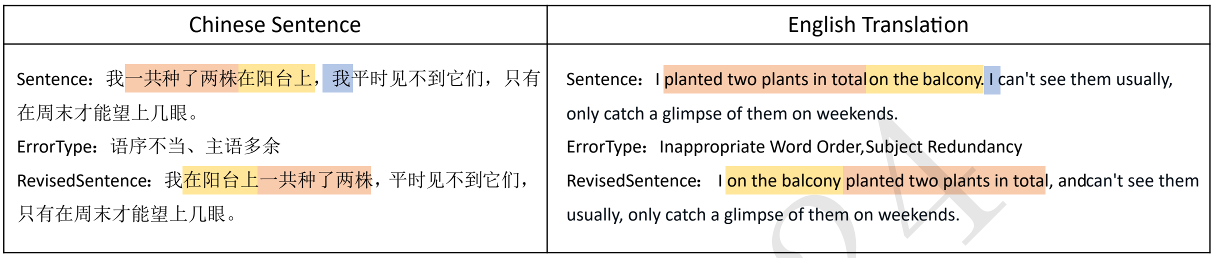
## 1 Introduction

Education is an enduring and evolving journey, continuously reshaping itself, particularly in the wake of the Internet's proliferation and the development of Large Language Models (LLMs) (Zhao et al., 2023), which has also significantly expanded the scope of Chinese essay evaluation. The marked increase in the volume of essays requiring evaluation has highlighted concerns regarding the cost-effectiveness and efficiency of manual essay corrections, positioning these issues as salient factors in contemporary educational methodologies. In light of these challenges, a growing number of scholars and educational institutions have initiated investigations into the feasibility of utilizing computer technologies for Automated Essay Correction (AEC) (Rudner et al., 2006; Ramesh and Sanampudi, 2022). These relative methods fulfill a twofold purpose. **Firstly**, it facilitates the provision of objective, precise, and timely feedback by analyzing various dimensions of an essay, such as language, content, and structure, and addressing inherent writing challenges. This, in turn, potentially enhances students' comprehension of their writing difficulties, thereby improving their overall writing competencies. **Secondly**, it enables educators to more accurately assess students' writing proficiency and offer more focused instructional support, furthering the educational advancement of students. In practical educational settings, a critical component that teachers assess during essay evaluation is the **fluency** of expression. This aspect reflects the essay's coherence and grammatical accuracy, offering insights into the writer's proficiency and ability to convey ideas effectively. Improving fluency is essential for enhancing the accuracy of essay evaluations and raising the authors' writing standards.

Nevertheless, the current evaluation of essay fluency at primary and secondary education levels faces significant challenges: **1) Lack of detailed criteria** Most existing assessments focus broadly on overall essay quality without delving deeply into fluency. There is a noticeable absence of systematic criteria, which hampers a thorough understanding and development of students' writing skills. **2) Limited interpretability** Previous studies often approach fluency as merely a scoring endeavor, providing only an aggregate score or rating. Alternatively, they treat it as a basic Grammatical Error Correction (GEC) task (Gong et al., 2021; Tsai et al., 2020). Such approaches predominantly target simple grammatical

[*]Corresponding author.

mistakes, reviewing them through simple corrections like additions, deletions, and modifications, which limits comprehensive feedback on the text's flow and structure. However, these methodologies typically overlook the analysis of specific grammatical error types and fail to specify the exact nature of the errors encountered. Providing students with detailed descriptions of error types and suggested corrections is beneficial as it enables them to recognize their mistakes, refine their essays, and prevent the recurrence of these errors in future writing. **3) Scarcity of authentic data from primary and secondary educational contexts** There is a notable shortage of public datasets for studying essay fluency among Chinese primary and secondary school students. Previous research relying on GEC often utilizes rule-based or inter-language datasets from learners of Chinese as a second language. However, the errors in compositions written by native Chinese students are more diverse and require a deeper understanding of complex grammatical structures. Figure 1 presents examples of sentences taken from essays written by primary and secondary school students. These excerpts showcase various errors along with recommended corrections. Typically, each sentence contains multiple types of errors that extend beyond simple spelling mistakes, illustrating the complexity of issues found in student writing.



| Chinese Sentence | English Translation |
|---|---|
| Sentence：我一共种了两株在阳台上，我平时见不到它们，只有在周末才能望上几眼。<br>ErrorType：语序不当、主语多余<br>RevisedSentence：我在阳台上一共种了两株，平时见不到它们，只有在周末才能望上几眼。 | Sentence：I planted two plants in totalon the balcony. I can't see them usually, only catch a glimpse of them on weekends.<br>ErrorType：Inappropriate Word Order,Subject Redundancy<br>RevisedSentence：I on the balcony planted two plants in total, andcan't see them usually, only catch a glimpse of them on weekends. |

Figure 1: An example of our task. In modern Chinese, adverbials are typically positioned between the subject and the predicate rather than at the end of the sentence, thereby leading to an 'Inappropriate Word Order' error. Moreover, in the first two short sentences, there is a problem of 'Subject Redundancy' where the subject 'I' is repeated unnecessarily.

These shortcomings in current methodologies highlight the need for a more detailed and nuanced approach that not only identifies fine-grained errors but also delivers specific, actionable feedback to students. It underscores the importance of utilizing composition data from the authentic writing contexts of primary and secondary school students to better address their specific learning needs. Motivated by this, we present the CCL24-Eval task [0]: ***C**hinese **E**ssay **F**luency **E**valuation* (**CEFE**), which aims to identify and correct errors that impede the fluency of writing in essays by primary and secondary school students and to assess the overall fluency level of an essay. Compared to Task 8 of the CCL23-Eval (Shen et al., 2023), we have introduced a new track and removed an original one, taking into account their practical application and relevance. This task featured three tracks: (1) *Coarse-grained and fine-grained error identification*; (2) *Error sentence rewriting*; (3) *Essay fluency level recognition*, aiming at providing a higher-quality evaluation of fluency in primary and secondary school essays.

This task attracted 29 teams to sign up for the competition, and in the end, we received 180 submissions from 10 teams. The task description is presented in Section 2. We describe the data we used in this task in Section 3. We explain baselines used for each track and list participants' information and results from their submissions and provide a more in-depth discussion in Section 5.

## 2 Task Description

Our evaluation is structured into three distinct tracks, each crafted to tackle specific aspects of identifying and correcting errors in essays written by primary and secondary school students. These tasks are designed to shed light on the common types of errors these students make, offering a basis for focused enhancements in their writing skills.

[0] https://github.com/cubenlp/2024CCL_CEFE

## 2.1 Track 1: Coarse-Grained and Fine-Grained Error Identification

Track 1 is dedicated to identifying types of grammatical errors in compositions from primary and secondary school students. Traditional methods often overlook these errors, failing to specifically highlight the diverse writing challenges students face. This track addresses the issue through two analytical lenses: character-level and component-level errors. We define four broad categories of grammatical errors: *Character-Level Error **(CL)***, *Incomplete Component Error **(IC)***, *Redundant Component Error **(RC)***, and *Incorrect Constituent Combination Error **(ICC)***. Additionally, we have outlined fourteen fine-grained error types, offering a deeper insight into the potential mistakes in student writing. Given the developmental stage of primary and middle school students, their compositions often contain multiple errors within the same sentence, making this task particularly challenging. Consequently, this track is structured as a multi-label classification task. Overall, Track 1 includes a total of 4 coarse-grained error types and 14 fine-grained error types. Detailed descriptions and examples of each error type are provided on the competition homepage, and the specific category definitions are as follows:

**Character-Level Error (CL)** includes four fine-grained error types: Word Missing (**WM**), where a word in a commonly used fixed collocation is missing from the sentence and needs to be added; Typographical Error (**TE**), where there are typos in the sentence that need to be revised or deleted; Missing Punctuation (**MP**), where punctuation is missing from the sentence and needs to be added; and Wrong Punctuation (**WP**), where the punctuation used in the sentence is wrong and needs to be revised or deleted.

**Redundant Component Error (RC)** is composed of three fine-grained error types: Subject Redundancy (**SR**), which occurs when a complex adverb is followed by a repeated subject referring to the same entity, and the modification is to delete one subject; Particle Redundancy (**PR**) refers to the redundant use of particles, which should be deleted during editing; Other Redundancy (**OR**) refers to any redundant elements not covered by the previous types, which should also be deleted in modification.

**Incomplete Component Error (IC)** consists of four fine-grained error types: Unknown Subject (**US**), which occurs when the sentence lacks a subject or the subject is unclear, and the solution is to add or clarify the subject; Predicate Missing (**PM**) refers to a sentence lacking verbs, which may be corrected by adding predicates; Object Missing (**OBM**) means that a sentence lacks an object, and the solution is to add an object; Other Missing (**OTM**) refers to other missing components besides the incomplete subject, predicate, and object, which may be corrected by adding the missing components except for the subject, predicate, and object.

**Incorrect Constituent Combination Error (ICC)** includes three fine-grained error types: Inappropriate Verb-Object Collocation (**IVOC**) refers to the predicate and object not being properly matched, and may be corrected by replacing either the predicate or object with other words; Inappropriate Word Order (**IWO**) means that the order of words or clauses in the sentence is unreasonable, and may be corrected by rearranging some words or clauses; Inappropriate Other Collocation (**IOC**) refers to any element in the sentence not covered by the previous types being improperly matched, and may be corrected by replacing it with other words.

## 2.2 Track 2: Error Sentence Rewriting

Track 2 focuses on the rewriting of incorrect sentences in compositions by primary and secondary school students. The main challenge of this track is to devise a minimal modification strategy for these erroneous sentences, ensuring that the original meaning is preserved. The corrections should involve as few changes as necessary because over-modifying can obscure the original errors, making it difficult for students to recognize and learn from their mistakes. This is crucial for teachers to better understand the writing challenges their students face, and to aid in improving their writing skills. It emphasizes the importance of maintaining the student's original thought process while steering them towards grammatical accuracy and clearer expression.

### 2.3 Track 3: Essay Fluency Level Recognition

Track 3 is designed to assess the overall fluency of an entire essay by examining the organization of words, sentences, and paragraphs. By evaluating the fluency of essays, this approach offers teachers a more efficient and intuitive method to assess students' writing skills. It also provides students with a clearer understanding of their writing performance. Defined as a multi-label classification task, Track 3 categorizes essays into three fluency levels: *excellent*, *average*, and *failing*. This classification helps in distinguishing essays based on their coherence and structural organization, allowing for targeted feedback that can guide improvements in students' writing abilities.

## 3 Datasets

In an effort to enhance research on essay fluency among primary and secondary school students, we meticulously annotated a dataset with fine-grained grammatical error types and provided corresponding corrections that impact sentence fluency. We developed the second **C**hinese **E**ssay **F**luency **E**valuation (**CEFE 2.0**) dataset based on CEFE 1.0 dataset (Shen et al., 2023), aiming to offer in-depth insights into the typical grammatical errors students make. This fine-grained dataset not only helps in identifying common mistakes but also supports the development of targeted teaching strategies to improve writing skills in young learners.

### 3.1 Data Collection

The foundational material for our dataset was derived from actual essays written by primary and secondary school students during their examinations. These compositions span a variety of genres, including character and scene descriptions, chosen for their genuine representation of students' writing skills. This choice of data source is pivotal as it captures the authentic richness of real-world writing scenarios. Exam essays particularly offer unfiltered insights into the writing abilities, habitual patterns, and prevalent errors among students in these educational stages. The diversity and complexity of the errors and required revisions found in these essays mirror the real challenges students face, providing a robust basis for our research. By utilizing these authentic compositions, our findings and proposed solutions remain highly relevant and directly applicable to improving student writing, thereby maximizing the impact of our work.

### 3.2 Data Annotation

The annotation team was composed of four undergraduate students, four postgraduate students specializing in language-related disciplines, and four expert reviewers with backgrounds in Chinese teaching. This diverse group was responsible for identifying error types and suggesting sentence revisions, adhering to the principle of **minimal changes**. Prior to beginning their tasks, all annotators underwent a training session designed to familiarize them with the specific annotation guidelines. The annotation process was structured as follows: an initial annotation was conducted collaboratively by an undergraduate and a postgraduate student. Subsequently, expert reviewers performed a verification pass to confirm the accuracy and reliability of the annotations, making necessary adjustments where needed. The annotated data was organized into five groups, and the team conducted weekly online meetings to discuss prevalent issues and refine their approach. This comprehensive process not only focused on pinpointing specific errors but also on providing actionable correction suggestions. Such a dual approach enhances the clarity of the feedback and equips students with the tools they need to improve their writing skills effectively.

### 3.3 Data Statistics

This section delineates the distribution of training, validation, and test datasets for each track. Given the prevalent scarcity of annotated data in real-world contexts, participants are tasked with developing robust models for assessing sentence fluency using a limited dataset. The test dataset includes both correct and intentionally flawed sentences, with a portion of the data reserved for blind evaluation. The statistics for our dataset are presented in Table 1.

|  | Train Set | Dev Set | Test Set |
|---|---|---|---|
| Track 1 | $1,000$ | $100$ | $2,000$ |
| Track 2 | $1,000$ | $100$ | $2,000$ |
| Track 3 | $100$ | $10$ | $2,000$ |

Table 1: The statistics of the **CEFE 2.0** dataset. The number for Track 1 and Track 2 corresponds to individual sentences, whereas for Track 3, it represents entire essays.

## 4 Evaluation Metrics

Different evaluation metrics are utilized across the various tracks of the task; however, the calculations for precision and recall remain consistent throughout. **Precision** is calculated as the ratio of correctly identified instances to the total number of instances identified by the model. Conversely, **Recall** is calculated as the ratio of correctly identified instances to the total number of instances labeled in the ground truth. The **F1-score**, frequently employed in binary or multi-class classification tasks, represents the harmonic mean of precision and recall, and is computed using the formula: $F_1 = \frac{2PR}{P+R}$.

### 4.1 Track1: Coarse-Grained and Fine-Grained Error Identification

The total score of Track1 is composed of two parts: coarse-grained and fine-grained wrong sentence identification score. The specific calculation method is as follows:

$$Score_{Track1} = 0.5 * F_1^{Coarse-grained} + 0.5 * F_1^{Fine-grained} \tag{1}$$

Specifically, precision (P), recall (R), and micro $F_1$ are used to evaluate the recognition effect of coarse and fine-grained wrong sentence types (See details in Section 2.1).

### 4.2 Track2: Error Sentence Rewriting

Due to the variety of rewriting outcomes, we assess the results of the model from two distinct perspectives:

**Comparison with Gold References:** We utilize three evaluation metrics: 1) **Exact Match (EM)**: This metric calculates the percentage of sentences generated by the model that perfectly align with the gold standard references. 2) **Edit Metrics** (proposed by MuCGEC) (Zhang et al., 2022): This method transforms error-correct sentence pairs into a series of operations and compares these operations produced by the model against the correct references, subsequently calculating precision, recall, and $F_{0.5}$ scores. 3) **BLEU** (Papineni et al., 2002): This metric assesses the N-gram overlap between sentences generated by the model and the correct references, providing a measure of linguistic similarity.

**Correctness and Reasonableness of Results:** We also apply three metrics to evaluate the rewritten sentences: 1) **Perplexity (PPL)**: Utilizing BERT (version bert-base-chinese) (Kenton and Toutanova, 2019), this metric gauges the fluency and predictability of the rewritten sentences. 2) **Levenshtein Distance**: This measures the edit distance between the rewritten sentence and the original, aiming to achieve accurate corrections with minimal edits to maintain clarity in understanding the nature of the errors. 3) **BERTScore** (Zhang et al., 2020): This score quantifies the semantic similarity between the rewritten and original sentences, ensuring the corrections maintain contextual integrity.

These metrics are subsequently weighted to compute a final score, effectively balancing various aspects of quality in the rewritten sentences:

$$Score_{Track2} = (EM + BLEU + F_{0.5} + BERTScore)/4 - Levenshtein - PPL_{BERT} \tag{2}$$

### 4.3 Track 3: Essay Fluency Level Recognition

To evaluate the classification performance of elementary and secondary school essay fluency ratings, we employ a range of metrics: Accuracy (Acc), Precision (P), Recall (R), Macro F1, and Quadratic

Weighted Kappa (QWK). Given that the QWK ranges from [-1, 1], we first normalize it to the interval [0, 1] before incorporating it into the weighted final score calculation. The composite score for track three is calculated as follows:

$$Score_{Track3} = 0.5 * F_1 + 0.2 * QWK + 0.3 * Acc \tag{3}$$

## 5 Results and Analysis

### 5.1 Baselines

We provide the results of our baseline models as a reference. For Track 1 and Track 3, we fine-tuned BERT (version *bert-base-chinese*) (Kenton and Toutanova, 2019) over the corresponding training datasets for 5 epochs, utilizing batch sizes ranging from 16 to 24, a learning rate of $2 \times 10^{-5}$, and employing the Adam optimizer. For Track 2, we fine-tuned BART (version *bart-base-chinese*) (Lewis et al., 2020) on the training dataset for 5 epochs, with a fixed batch size of 16, a learning rate of $2 \times 10^{-5}$, and the AdamW optimizer. Detailed results of these baseline models are provided in Section 5.

### 5.2 Results

In our competition, a total of 10 teams submitted their final results. The basic information about them are detailed in Table 2. Table 3 displays the final results of the participating teams in this competition. The average score across the three tracks will serve as the final score for the current team.

| ID | Team Name | Organization | Track 1 | Track 2 | Track 3 |
|---|---|---|---|---|---|
| 1 | Smartdot | Smartdot Technologies Co.,Ltd. | ✓ | ✓ | ✓ |
| 2 | BNU | Beijing Normal University | ✓ | ✓ | ✓ |
| 3 | ZUT-POLab | Zhongyuan University of Technology | ✓ | ✓ | ✓ |
| 4 | GDUFS-CL | Guangdong University of Foreign Studies | ✓ | ✓ | ✓ |
| 5 | SIAT-UI | Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences | ✓ | ✓ | ✓ |
| 6 | ZZU1 | Zhengzhou University | ✗ | ✓ | ✗ |
| 7 | CPIC | China Pacific Insurance(Group) Co., Ltd. | ✓ | ✓ | ✓ |
| 8 | KUST | Kunming University of Science and Technology | ✓ | ✓ | ✓ |
| 9 | DeakinAI | Deakin University | ✓ | ✓ | ✓ |
| 10 | ZZU2 | Zhengzhou University | ✓ | ✓ | ✗ |
| Total Number | | 29 | 27 | 28 | 26 |

Table 2: The basic information of the participants with a total of 29 teams, where 27 teams for Track 1, 28 teams for Track 2 and 26 teams for Track 3.

## 6 Participant Systems

In this task, the competing teams employed a variety of approaches to detect and correct errors in student essays and to assign grades based on fluency. This section will provide an overview of the methods that were successful in each track. The unique approaches of each team demonstrate the diversity of methods that can be exploited in automated essay scoring and present various potential directions for future research.

### 6.1 Track1: Coarse-Grained and Fine-Grained Error Identification

For Track 1, ***Smartdot*** implemented a two-stage fine-tuning strategy, initially utilizing publicly available datasets to augment data through grammar error substitutions and fine-tuning using the expanded dataset with the UTC model (Universal Text Classification). Subsequently, they leveraged the training data provided by the competition for a second phase of fine-tuning. ***ZUT-POLab*** first augmented the training data through operations such as insertion, substitution, and deletion, and then fine-tuned the ERNIE 1.0 model using the expanded dataset. ***GDUFS-CL*** analyzed two fine-grained errors, utilizeda binary

| Rank | Team Name | Track 1 | Track 2 | Track 3 | Score |
|---|---|---|---|---|---|
| 1 | Smartdot | 48.48 | 52.41 | 50.56 | 50.48 |
| 2 | BNU | 45.19 | 43.90 | 45.84 | 44.98 |
| 3 | ZUT-POLab | 41.66 | 43.49 | 46.98 | 44.04 |
| 4 | GDUFS-CL | 36.47 | 41.09 | 51.96 | 43.17 |
| 5 | SIAT-UI | 37.26 | 42.48 | 47.64 | 42.46 |
| 6 | ZZU1 | - | 46.06 | - | 41.72 |
| 7 | CPIC | 35.42 | 45.32 | 33.38 | 38.04 |
|  | **baseline** | 34.62 | 30.00 | 44.48 | 36.37 |
| 8 | KUST | 34.19 | 24.87 | 45.71 | 34.92 |
| 9 | DeakinAI | 0.00 | 16.77 | 45.40 | 20.72 |
| 10 | ZZU2 | 0.00 | 0.00 | - | 14.83 |

Table 3: Final scores of the participating teams. ”-” indicates that the team did not submit evaluation results on the track, and the overall score is calculated based on the baseline.

classification model for prediction optimization, compared and selected trainingcorpora, and trained a coarse-grained model based on the Chinese Learner 4W corpus.

Some teams also leveraged Large Language Models (LLMs) to assist in their tasks. ***BNU*** first undertook data augmentation, scraping essays from websites such as primary and secondary school essay networks. They then used LLMs like GPT-3.5 to generate specific types of flawed sentences starting from given correct sentences. After constructing the data, they fine-tuned BERT using both the augmented and provided training data. ***SIAT-UI*** utilized the released task data to fine-tune the Qwen1.5-7b-chat model.utilized the released task data to fine-tune the Qwen1.5-7b-chat model.

### 6.2 Track2: Error Sentence Rewriting

For Track 2, ***Smartdot*** proposed a two-stage strategy: in the first stage, they pre-trained BART using pseudo-native data and the NaSGEC dataset, and incorporated SynGEC for grammatical error correction. In the second stage, they fine-tuned the model on the training dataset provided by the competition. ***ZZU1***, based on the BART model, proposed employing multi-pass decoding within the sequence-to-sequence framework to iteratively refine the corrections from the previous round, and additionally introduced an early stopping mechanism to reduce computational costs. ***ZUT-POLab*** proposed a diffusion generative model where, during the forward process, each step's text is encoded using ERNIE 1.0, and in the reverse process, the text modeling capabilities of ERNIE are utilized to progressively decode the masked tokens.

***BNU*** utilized LLMs to address this task. They initially constructed hybrid training data by enhancing the quality of the data created for Track 1 through data processing and cleaning efforts, and by integrating open-source datasets such as MuCGEC. Subsequently, they fine-tuned the Qwen-14B model based on the crafted hybrid data and designed specific prompt templates to reinforce the model's generalization ability.

### 6.3 Track 3: Essay Fluency Level Recognition

For Track 3, ***GDUFS-CL*** employed back-translation techniques to construct pseudo-data with triple-labeled fluency ratings for pre-training and adapting an NSP-based strategy to effectively utilize contextual information and avoid long sequence dependencies. ***Smartdot*** The Smart team initially selected the TextRCNN-NeZha model as their foundational model. They then introduced the MulDrop strategy and employed the DCE loss for the classification of essay fluency levels. ***SIAT-UI*** utilized the data provided by the competition to fine-tune the Bert-Large-Chinese model. ***ZUT-POLab*** first employed operations such as insertion, substitution, and deletion to augment the training data, and then fine-tuned the ERNIE 3.0 model using the expanded dataset. ***BNU*** employed a multitask learning approach, where, in addition to the primary task of essay fluency level recognition, they also trained an auxiliary model for the identification of grammatical errors in these essays to support the fluency grading.

## 7 Conclusions and Future Work

This paper presents an overview of the CCL24-Eval Task the second ***C**hinese **E**ssay **F**luency **E**valuation* (**CEFE**). We conduct this evaluation using our meticulously annotated **CEFE 2.0** dataset. The evaluation is divided into three distinct tracks: (1) Coarse-grained and fine-grained error identification; (2) Error sentence rewriting; and (3) Essay Fluency Level Recognition. We received a total of 29 completed registration forms, culminating in 180 submissions from 10 participating teams. In addition, we provide a comprehensive analysis and summary of the methodologies employed by the participants, which will contribute to future research in this field of natural language processing. The findings indicate that the employment of LLMs and the application of data augmentation techniques contribute to enhancing the aggregate scores.

## Acknowledgements

We appreciate the support from National Natural Science Foundation of China with the Main Research Project on Machine Behavior and Human Machine Collaborated Decision Making Methodology (72192820 & 72192824), Pudong New Area Science Technology Development Fund (PKX2021-R05), Science and Technology Commission of Shanghai Municipality (22DZ2229004) and Shanghai Trusted Industry Internet Software Collaborative Innovation Center.

## References

Jiefu Gong, Xiao Hu, Wei Song, Ruiji Fu, Zhichao Sheng, Bo Zhu, Shijin Wang, and Ting Liu. 2021. Iflyea: A chinese essay assessment system with automated rating, review generation, and recommendation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 240–248.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

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

Lawrence M Rudner, Veronica Garcia, and Catherine Welch. 2006. An evaluation of intellimetric™ essay scoring system. *The Journal of Technology, Learning and Assessment*, 4(4).

Xinshu Shen, Hongyi Wu, Xiaopeng Bai, Yuanbin Wu, Aimin Zhou, Shaoguang Mao, Tao Ge, and Yan Xia. 2023. Overview of CCL23-eval task 8: Chinese essay fluency evaluation (CEFE) task. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations)*, pages 282–292, Harbin, China, August. Chinese Information Processing Society of China.

Chung-Ting Tsai, Jhih-Jie Chen, Ching-Yu Yang, and Jason S Chang. 2020. Lingglewrite: a coaching system for essay writing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 127–133.

Tianyi Zhang, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yue Zhang, Zhenghua Li, Zuyi Bao, Jiacheng Li, Bo Zhang, Chen Li, Fei Huang, and Min Zhang. 2022. MuCGEC: a multi-reference multi-source evaluation dataset for Chinese grammatical error correction. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3118–3130, Seattle, United States, July. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.