Rethinking Overcorrection for Grammatical Error Correction Evaluation

Anonymous ACL submission

Abstract

Grammatical Error Correction (GEC) requires objective and effective evaluation metrics that align with human judgments. Current metrics, whether reference-based or reference-less, suffer from biases and fail to fully capture key principles like grammaticality, faithfulness, and fluency. In this paper, we propose a **W**eighted **O**vercorrection-**D**ecoupled **A**nd **F**luency **E**valuation metric, **WODAFE**, which integrates reference-based and reference-less evaluation through two components: (1) a generalized F-score that decouples overcorrection for edit-level evaluation, and (2) a fluency score for sentence-level assessment. WODAFE achieves state-of-the-art alignment with human judgments in specific levels, especially Spearman $\rho = 1$ at edit-level. Our analysis highlights the potential of overcorrection weight tuning to reduce biases of failing to cover all references in reference-based metrics. We also apply our metric to the CoNLL14 benchmark and reveal key evaluation differences.

1 Introduction

Grammatical Error Correction (GEC) refers to the automatic detection and rectification of linguistic errors in text (Bryant et al., 2023). A primary challenge in GEC evaluation is the development of objective, reliable metrics that accurately reflect human judgment and preferences.

Automatic evaluation of GEC categorize into *reference-based* and *reference-less* methods (Maeda et al., 2022). Although recent studies aim to develop metrics in both categories that closely align with human judgments, significant limitations remain. *Reference-based* metrics face challenges in preparing exhaustive correction references (Choshen and Abend, 2018b), while *reference-less* metrics, which evaluate text using generative models, exhibit inherent biases (Deutsch et al., 2022). Consequently, most existing metrics skew toward specific biases, failing to fully address the three principles essential for GEC evaluation: grammaticality, faithfulness[1], and fluency. Although recent metrics achieve higher human correlation, they lack interpretability to some extent (Islam and Magnani, 2021; Kobayashi et al., 2024a). Additionally, while human evaluations emphasize granularity at both *edit-level* and *sentence-level* (Kobayashi et al., 2024b), few metrics incorporate this dual perspective.

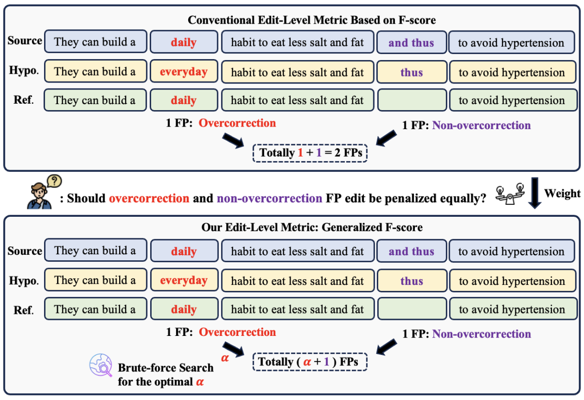


Figure 1: Comparison of conventional edit-level metric based on F-score and our generalized F-score. Conventional F-score treats Overcorrection and Non-overcorrection FP edit equally, while our task highlights the relative weight of overcorrection and fine-grained evaluation.

In GEC, *false positives* (FPs) manifest as either incorrect edits to valid text (FP_{noc}) or unnecessary edits (FP_{oc}), namely ***overcorrection***[2]. Large Language Models (LLMs) frequently exhibit overcorrection tendencies, yielding high recall but low precision (Loem et al., 2023). Our preliminary study in Section 2.1 confirms the prevalence of this issue. Though recent work (Ye et al., 2024) disentangles overcorrection for separate evaluation,

[1] Grammaticality requires error-free sentences; faithfulness ensures corrections preserve original meaning and syntax.

[2] FP_{noc} represents FP of non-overcorrection, while FP_{oc} means FP of overcorrection

the relative importance of FP_{oc} (overcorrection) versus FP_{noc} from a human judgment perspective remains unexplored.

To address these gaps, we propose **WODAFE**, Weighted Overcorrection-Decoupled And Fluency Evaluation, a hybrid reference-based and reference-less metric adhering to grammaticality, faithfulness, and fluency. Our approach operates at two levels: 1) Edit-level: **generalized F-score** $F_{\beta-G}$, combining **modified F-score** $F_{\beta-M}$ and an Overcorrection Score (**OCScore**), which decouples FP types and incorporates BERT-based weights to penalize edit errors discriminately. 2) Sentence-level: **fluency score** $f(x)$ assessing overall text coherence and fluency. By integrating these components with adjustable weights, our comprehensive metric achieves high correlation with human judgments (e.g. Spearman $\rho = 1$ at edit-level) and mitigates reference-based biases through overcorrection weight tuning.

Most existing benchmarks like CoNLL14 (Ng et al., 2014), BEA19 (Bryant et al., 2019) and JFLEG (Napoles et al., 2017) rely on conventional metrics like M^2 (Dahlmeier and Ng, 2012), ERRANT (Bryant et al., 2017), and GLEU (Napoles et al., 2015), which exhibit limited human correlation in the meta-evaluation experiment (Ye et al., 2023). While newer metrics improve alignment, their adoption on established benchmarks remains sparse. We bridge this gap by applying WODAFE to CoNLL14 benchmark, revealing critical discrepancies in evaluation outcomes.

Our contributions are threefold:

1. The first hybrid metric combining reference-based and reference-less methods, achieving state-of-the-art human judgment alignment (e.g., edit-level Spearman $\rho = 1$)
2. A novel framework for quantifying overcorrection's relative impact, demonstrating its utility in reducing reference-based metric biases.
3. Benchmark analysis exposing key differences between highly human-aligned metrics and conventional evaluations.

2 Method

2.1 Overcorrection

Overcorrection refers that the model unnecessarily changes parts of a sentence that are already grammatically correct. In GEC, general FP can be divided into two categories: 1) FP_{noc}: making incorrect edits to necessary positions, and 2) FP_{oc}: making unnecessary edits, namely overcorrection. Based on *CLEME*(Ye et al., 2023), an extension of the metric *ERRANT*, which merges overlapping edits into a chunk, overcorrection can be defined as the corrected hypothesis chunks while their corresponding reference chunks remain unchanged. With this definition, we preliminarily studied the frequency and severity of overcorrection in a subset of the CoNLL14 test set with 2 annotators(Ng et al., 2014), consisting of 391 sentences across 14 system outputs. As shown in Figure 2, overcorrection accounts for at least half of the FPs in each system, and the F-score will increase by at least 0.1 if overcorrection is absent from a system's output. This illustrates that overcorrection is ubiquitous, significant, and has a substantial impact on evaluation results. Therefore, it is essential to decouple FP_{oc} and FP_{noc} from general FP and study them separately.

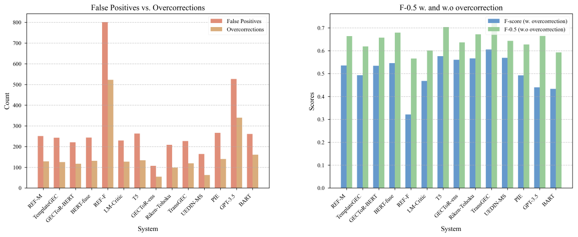


Figure 2: Preliminary study on overcorrection. The left image presents the frequency of FP versus overcorrection, while the right image shows the comparison between the F-score with and without overcorrection.

2.2 Edit-level Metric: Generalized Weighted F-score

To decouple FP_{noc} and FP_{oc} and obtain a fine-grained evaluation metric, we propose a generalized F_β score ($F_{\beta-G}$) based on $FP = FP_{noc} + FP_{oc}$. As shown in 1, F_β can be rewritten as a fractional equation comprising a **modified** F_β **score** ($F_{\beta-M}$) and an **Overcorrection score (OCScore)**. $F_{\beta-M}$ focuses solely on errors that need to be corrected, while OCScore measures the severity of overcorrection. It is worth mentioning that this disentanglement adheres to the two golden principles of minimal-edit GEC: grammaticality and faithfulness (Ye et al., 2024) . Grammaticality requires that all grammatical errors are accurately corrected (as measured by $F_{\beta-M}$), while faithfulness demands that the corrections preserve the original textual meaning and syntactic structure (as evaluated by the OCScore).

$$\frac{1}{F_\beta} = \frac{(1+\beta^2)\cdot TP + \beta^2 \cdot FN + FP_{noc}}{(1+\beta^2)\cdot TP} + \frac{FP_{oc}}{(1+\beta^2)\cdot TP} = \frac{1}{F_{\beta-M}} + \frac{1}{OCScore} \tag{1}$$

Since our edit-level metric is based on *CLEME*, an initial length weighting is added to the chunk evaluation, compensating for long chunk matching (Ye et al., 2023). Additionally, it is unreasonable to treat every edit equally when classifying them as true positives (TP), false positives (FP), and false negatives (FN), because not all edits are of equal importance. For example, correcting two minor errors (e.g., a definite article error related to the usage of "the" and "a") may not be as important as correcting a single but significant semantic error in terms of human preference. Therefore, we propose weighted edits based on BERTScore (Zhang et al., 2019), a pretraining (PT)-based (Gong et al., 2022) method that calculates similarity between two sentences and captures rich information in semantics, syntax, and authenticity.

$$X' = replace(X, e_{hyp}) \tag{2}$$

$$w = |PTScore(X', R) - PTScore(X, R)| \tag{3}$$

where the function replace() is intended to replace a specific chunk of the source X with the hypothesis chunk e_{hyp}. Here, R denotes the reference sentence. With this method, edits of varying importance can be rewarded or penalized appropriately. Furthermore, to quantify the difference between the overcorrection edits (FP_{oc}) and the necessary errors (FP_{noc}), we apply a factor α as an extra weight for FP_{oc}. The complete edit-level metric is shown in Equation 4.

2.3 Sentence-level Metric: Fluency

To compensate for the inherent biases of reference-based evaluation metrics (Choshen and Abend, 2018b), reference-less metric focusing on the overall sentence-level fluency is proposed. Carrying on the work of (Ge et al., 2018), we define fluency based on cross entropy.

$$f(x) = \frac{1}{1+H(x)},\ H(x) = -\frac{\sum_{i=1}^{|x|} logP(x_i|x_{<i})}{|x|} \tag{5}$$

where $P(x_i|x_{<i})$ is the probability of x_i given context $x_{<i}$, computed by a language model (gpt-2 in our work), and $|x|$ is the length of sentence x. Intuitively, the fluency score will be higher if the sentence is more grammatical and fluency because of higher probability.(Yasunaga et al., 2021)

2.4 Comprehensive Metric

Recent studies (Kobayashi et al., 2024c; Yoshimura et al., 2020) tend to combine different evaluation metric approaches because the complementary strengths of various metrics can be combined to achieve superior results. Consequently, we combined the two metrics described above to create a more comprehensive and robust metric by adjusting the weight. This metric combines reference-based and reference-less methods, allowing for a flexible edit-level and sentence-level score ratio adjustment in specific situations and domains.

$$Final\ Score = (1-\gamma)\cdot F_{\beta-G} + \gamma \cdot fluency \tag{6}$$

3 Experiments

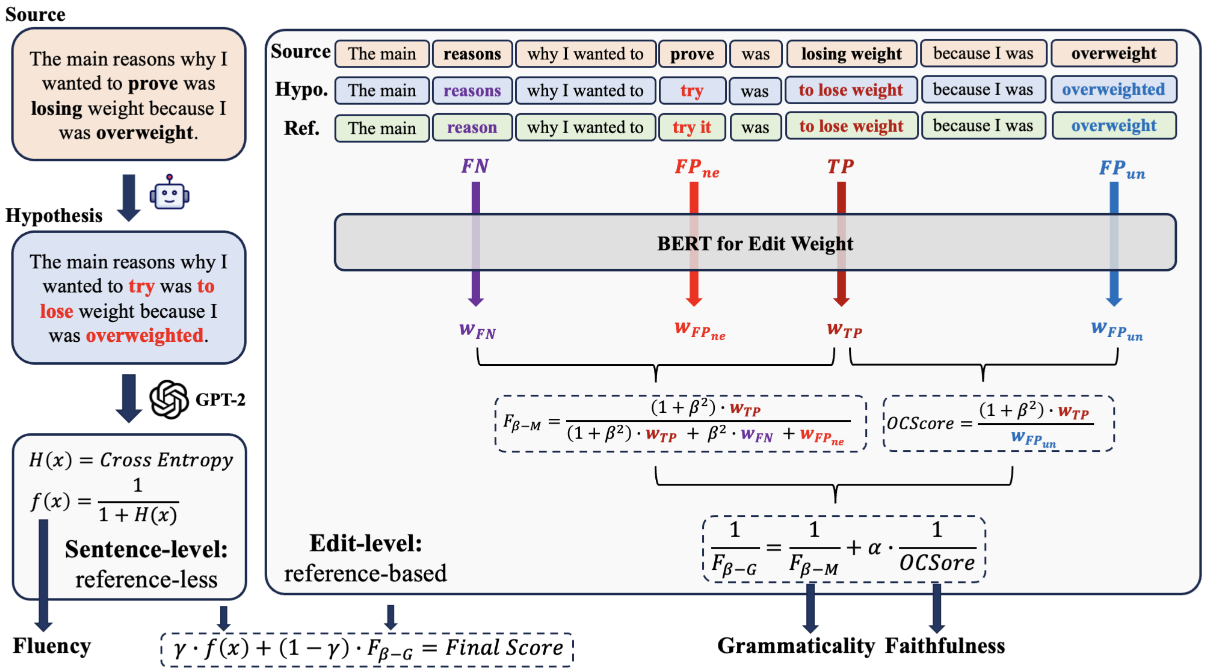
3.1 Meta-evaluation Setup

Human Evaluation Dataset. There are two mainstream human judgment datasets: GJG15(Grundkiewicz et al., 2015) and SEEDA(Kobayashi et al., 2024b). However, GJG15 was later found to be problematic, and many of the conclusions drawn using these datasets were called into question(Chollampatt and Ng, 2018; Choshen and Abend, 2018a). Additionally, GJG15 may yield different results depending on granularity, such as sentence-level or edit-level evaluations(Kobayashi et al., 2024b). SEEDA, on the other hand, performs human evaluations based on two different granularities: SEEDA-E for edit-based evaluation and SEEDA-S for sentence-based evaluation, making it more reliable. Therefore, we conduct comprehensive experiments using the SEEDA dataset.

Considered metrics. To align with the granularity, we use two types of evaluation metrics: Edit-Based Metrics (EBMs) and Sentence-Based Metric (SBMs) for comparison. EBMs includeds four metrics: M^2, ERRANT, GoToScore (Gotou et al., 2020) and PT-M^2(Gong et al., 2022). SBMs also uses four metrics: GLEU, Scribendi Score(Islam and Magnani, 2021), SOME (Yoshimura et al., 2020) and IMPARA (Maeda et al., 2022)

System-level Meta-evaluation. System-level correlation is computed by comparing the ranking of the participating systems by humans and the ranking generated by the metric scores. We utilize the system human scores derived from human rankings of systems using TrueSkill(TS) (Sakaguchi et al., 2014). System-level metrics compute the system score based on the whole

$$\frac{1}{F_{\beta-G}} = \frac{(1+\beta^2)\cdot w_{TP} + \beta^2 \cdot w_{FN} + w_{FP_{noc}}}{(1+\beta^2)\cdot w_{TP}} + \alpha \cdot \frac{w_{FP_{oc}}}{(1+\beta^2)\cdot w_{TP}} = \frac{1}{F_{\beta-M}} + \alpha \cdot \frac{1}{OCScore} \tag{4}$$



Source
The main reasons why I wanted to prove was losing weight because I was overweight.
Hypothesis
The main reasons why I wanted to try was to lose weight because I was overweighted.
GPT-2
$H(x) = Cross\ Entropy$
$f(x) = \frac{1}{1+H(x)}$
Sentence-level:
reference-less
Source | The main | reasons | why I wanted to | prove | was | losing weight | because I was | overweight
Hypo. | The main | reasons | why I wanted to | try | was | to lose weight | because I was | overweighted
Ref. | The main | reason | why I wanted to | try it | was | to lose weight | because I was | overweight
FN
FP_{ne}
TP
FP_{un}
BERT for Edit Weight
w_{FN}
$w_{FP_{ne}}$
w_{TP}
$w_{FP_{un}}$
$F_{\beta-M} = \frac{(1+\beta^2)\cdot w_{TP}}{(1+\beta^2)\cdot w_{TP} + \beta^2 \cdot w_{FN} + w_{FP_{ne}}}$
$OCScore = \frac{(1+\beta^2)\cdot w_{TP}}{w_{FP_{un}}}$
$\frac{1}{F_{\beta-G}} = \frac{1}{F_{\beta-M}} + \alpha \cdot \frac{1}{OCScore}$
Edit-level:
reference-based
Fluency
$\gamma \cdot f(x) + (1-\gamma)\cdot F_{\beta-G} = Final\ Score$
Grammaticality
Faithfulness

Figure 3: Overview of our approach **WODAFE**. $F_{\beta-M}$, $F_{\beta-G}$ and OCScore respectively denote modified F-score, generalized F-score and Overcorrection Score. There are two core modules in our approach: **edit-level** and **sentence-level** evaluation. The final score is derived by weighting the two scores.

corpus. For our metric, $F_{\beta-G}$ is computed in terms of the whole corpus. The fluency score, however, is computed by averaging the fluency scores of each sentence in the corpus. This is because directly utilizing the fluency score of the whole corpus is coarse-grained and unreliable. Finally, We measure both Pearson(r) and Spearman(ρ) correlation coefficients between our metric rankings and human rankings.

Sentence-level Meta-evaluation. In the sentence-level meta-evaluation, we use pairwise judgments from SEEDA. Our metric calculates average sentence-level scores for both $F_{\beta-G}$ and fluency. We employ Accuracy (Acc) and Kendall's rank correlation (τ) to investigate the correlation between human evaluations and metric scores.

3.2 Results

For our metric, we take the highest correlation coefficients as the final results by brute-force searching all possible values of α and γ. Table 1 reports the correlations between selected metrics and human judgments. In the system-level correlation of SEEDA-E, our metric outperforms all existing metrics, especially the Spearman correlation ρ=1.0 in the BASE[3] meta-evaluation, which has never been achieved before, to our knowledge. In the sentence-level correlation, our metric outperforms all existing metrics, whether EBMs or SBMs, and achieves state-of-the-art (SOTA) performance.

In terms of **the number of annotators** for the test set reference, we find that:

At the system level, the correlation coefficients remains essentially unchanged, except for the Spearman coefficient of SEEDA-S, which decreases to some extent.

At the sentence-level, utilizing a test set annotated by ten annotators yields higher correlations with human judgments compared to a test set annotated by only two annotators. We infer that more references help mitigate the inherent biases of the reference-based metric $F_{\beta-G}$, leading to a

[3] "BASE" uses the 12 systems excluding outliers, and "+ Fluent corr." adds two fluent corrected sentence system in ranking comparisons additionally.

Metric	System-level								Sentence-level							
	SEEDA-E				SEEDA-S				SEEDA-E				SEEDA-S			
	Base		+ Fluent corr.		Base		+ Fluent corr.		Base		+ Fluent corr.		Base		+ Fluent corr.	
	r	ρ	r	ρ	r	ρ	r	ρ	Acc	τ	Acc	τ	Acc	τ	Acc	τ
M^2	0.791	0.764	-0.239	0.161	0.658	0.487	-0.336	-0.013	0.582	0.328	0.527	0.216	0.512	0.200	0.496	0.170
ERRANT	0.697	0.671	-0.502	0.051	0.557	0.406	-0.587	-0.116	0.573	0.310	0.511	0.188	0.498	0.189	0.471	0.129
GoToScorer	0.901	0.937	0.667	0.916	**0.929**	0.881	0.627	0.881	0.521	0.042	0.505	0.009	0.477	-0.046	0.504	0.009
PT-M^2	0.896	0.909	-0.083	0.442	0.845	0.769	-0.162	0.336	0.587	0.293	0.542	0.200	0.527	0.204	0.528	0.180
GLEU	0.911	0.897	0.053	0.482	0.847	0.886	-0.039	0.475	0.695	0.404	0.630	0.266	0.673	0.351	0.611	0.227
Scribendi Score	0.830	0.848	0.721	0.847	0.631	0.641	0.611	0.717	0.377	-0.196	0.359	-0.240	0.354	-0.238	0.345	-0.264
SOME	0.901	0.951	0.943	0.969	0.892	0.867	0.931	0.916	0.747	0.512	0.743	0.494	0.768	0.555	0.760	0.531
IMPARA	0.889	0.944	0.935	0.965	0.911	0.874	**0.932**	**0.921**	0.742	0.502	0.725	0.455	0.761	0.540	0.742	0.496
Ours (two-annotators)	**0.931**	**1.0**	**0.962**	**0.987**	0.886	0.895	0.923	0.895	0.789	0.577	0.772	0.543	0.788	0.575	0.760	0.520
Ours (ten-annotators)	0.925	0.993	**0.962**	0.978	0.861	0.853	0.923	0.873	**0.799**	**0.598**	**0.780**	**0.559**	**0.789**	**0.578**	**0.768**	**0.535**

Table 1: Results of system-level and sentence-level meta-evaluations in SEEDA dataset. Two(Ten) annotators means the system outputs are evaluated on CoNLL test set annotated by two(ten) annotators. **Bold** indicates the **highest** correlation score and the underline indicates the second highest correlation score. The results of existing metrics are from (Kobayashi et al., 2024a)

Metric	System-level								Sentence-level							
	SEEDA-E				SEEDA-S				SEEDA-E				SEEDA-S			
	Base		+ Fluent corr.		Base		+ Fluent corr.		Base		+ Fluent corr.		Base		+ Fluent corr.	
	r	ρ	r	ρ	r	ρ	r	ρ	Acc	τ	Acc	τ	Acc	τ	Acc	τ
α (two-annotators)	0.120	0.195	0.000	0.035	0.005	0.085	0.000	0.035	0.480	0.480	0.460	0.460	0.155	0.155	0.235	0.235
γ (two-annotators)	0.865	0.825	0.005	0.885	0.015	0.77	0.075	0.895	0.985	0.985	0.985	0.985	0.955	0.955	0.965	0.965
α (ten-annotators)	0.140	0.005	0.000	0.005	0.005	0.005	0.000	0.005	0.740	0.740	0.460	0.460	0.065	0.065	0.355	0.355
γ (ten-annotators)	0.800	0.215	0.005	0.870	0.010	0.010	0.005	0.870	0.815	0.815	0.945	0.945	0.945	0.945	0.945	0.945

Table 2: Corresponding α and γ values when the maximum coefficient values are taken.

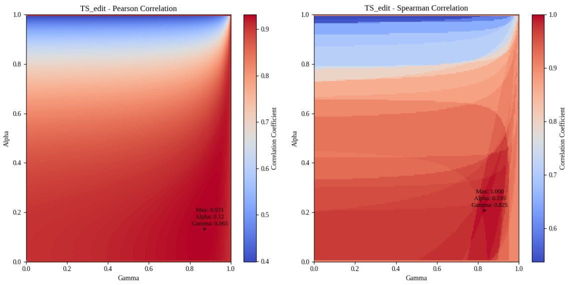


Figure 4: Heat map of system-level SEEDA-E (Two annotators). *TS_edit* denotes human scores are derived with TrueSKill(TS) approach and the meta-evaluation is taken in edit-level BASE setting.

final score that correlates more closely with human preferences on average.

Furthermore, we plot heat maps to study when our metric achieves the maximum correlation and to examine the relative importance of FP_{noc} and FP_{oc}, $F_{\beta-G}$ and $f(x)$ at the edit and sentence levels. Figure 4 presents the heat map for the system-level SEEDA-E (two annotators). Complete heat maps are provided in Appendix E. We find that:

At the system level, overcorrection (FP_{oc}) is much less important than FP_{noc} because the correlations increase rapidly as α decreases, peaking at $\alpha \in (0, 0.2)$. However, sentence-level fluency $f(x)$ is relatively more important than edit-level $F_{\beta-G}$, as correlations tend to be higher as γ increases, usually peaking at $\gamma \in (0.75, 0.9)$. This tendency is stronger and more significant in the meta-evaluation "+Fluent Corr", which means the fluency score $f(x)$ should be weighted more if we require to emphasize the fluency quality in GEC evaluation.

At the sentence level, on the one hand, overcorrection (FP_{oc}) is still relatively less important than FP_{noc} because the correlations increase as α decreases, even though with at a slower rate compared to that of system level, typically reaching the maximum value when $\alpha \in (0, 0.2)$ or around 0.5. On the other hand, sentence-level fluency weighs far more than edit-level $F_{\beta-G}$, with maximum correlation usually occurring when γ is near 0.9.

4 Analysis

4.1 Overcorretion

We observe that the number of overcorrection edits in a system is inversely correlated with the number of references (i.e., annotators). As shown in Figure 5, the number of FPs decreases by over one-third when annotators increase from two to ten. This reduction primarily stems from fewer overcorrections, as the quantitative difference between FPs aligns closely with the decline in overcorrection. Notably, edits previously classified as overcorrections now approximate the count of "new" FPs derived from 10-annotator references, reflecting the expanded reference coverage. This suggests that a large number of overcorrections are **inflation indicators** caused by insufficient reference coverage, a known bias in reference-based metrics (Choshen and Abend, 2018b), especically when the references are relatively few.

We further analyze the maximum[4] weight α required for the Spearman coefficient ρ to reach 0.95 in the *TS_edit* BASE setting. In the two-annotator configuration, ρ first achieves 0.95 when α decreases to 0.505, whereas the ten-annotator setting only requires α=0.995. This implies that comparable human correlations can be achieved even with insufficient and fewer references by lowering the relative weight α for overcorrection moderately, thus mitigating inherent biases of reference-based metrics. Adjusting α for overcorrection thus offers a viable method to enhance metric robustness.

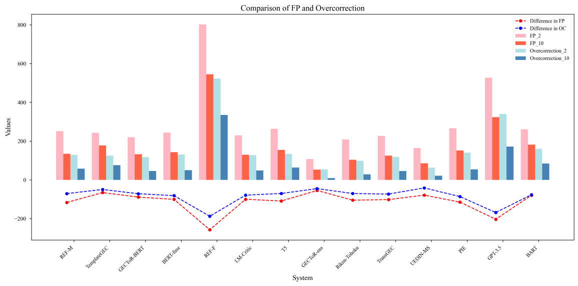


Figure 5: The comparison of FP and overcorrection when the test set is annotated by two (FP_2 and Overcorrection_2) and ten annotators (FP_10 and Overcorrection_10).

4.2 Ablation Study

Effect of decoupling FP. Here we demonstrate the effect of relative weight α by evaluating the test set annotated by two-annotators with generalized F-score $F_{\beta-G}$ that removes edit weights. Figure 6 shows that correlation coefficients decrease as α becomes larger and there is a huge gap between w. and w.o. decoupling. Besides, the correlation in other granularity like SEEDA-S also has the same tendency, demonstrating the relatively less importance of overcorrection compared to FP_{noc} and the effectivenss of our decoupling.

Effect of fluency score. Table 4 in appendix presents the correlation coefficients between human judgments and fluency metric in all granularities. We find that the fluency metric itself has significantly higher correlations with human judgments compared to mainstream and conventional metrics such as M^2, ERRANT and GLEU in specific granularities.

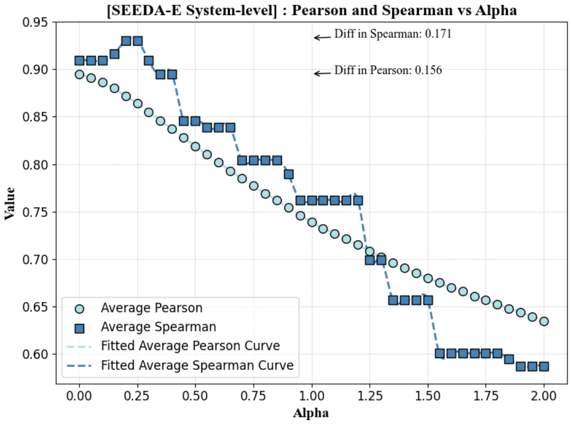


Figure 6: Pearson and Spearman coefficients between human judgments and generalized F-score $F_{\beta-G}$ with no edit weights in SEEDA-E. The "Diff" in the image is the difference between the maximum coefficient when overcorrection weight is adjustable and coefficient without decoupling FP (i.e. $\alpha = 1$)

4.3 Window Analysis

Window analysis (Kobayashi et al., 2024b) on the SEEDA dataset is performed by selecting a specific number of consecutive systems from the human rankings of the 12 systems and taking meta-evaluation in the selected systems in "BASE" configuration. Figure 7 shows the window analysis at both the edit-level and sentence-level in meta-evaluation, with a window size of four. We find that our metric, WODAFE, is exceptionally stable and robust across all window ranges for edit-level meta-evaluation. In sentence-level window analysis, WODAFE remains quite stable in higher-ranking ranges but suffers from low correlation when the ranking range is from 7th to 10th, similar to other metrics. Therefore, one of our future works is to improve the human correlations and stability of sentence-level evaluation.

4.4 Why not LLM?

Prior work (Kobayashi et al., 2024a) has observed the huge potential for LLM as a scorer in GEC task and get quite high correlations with human judgments. However, we find that utilizing LLM (e.g. GPT-4) as a scorer has these following shortcomings.

Lack of interpretability. Acting as "black-box"

[4] We study the "maximum" because the correlation coefficient increases as α decreases, as the heat map shows.

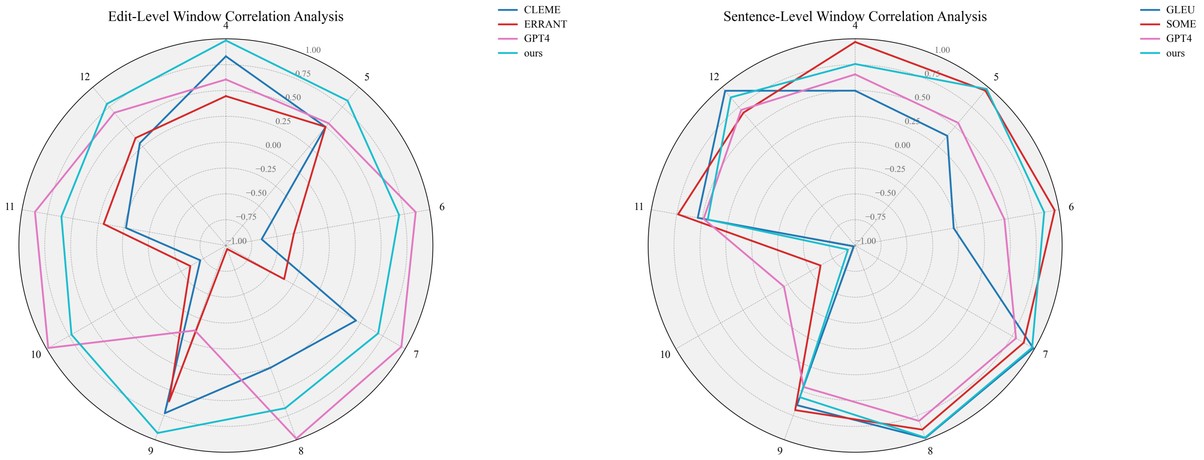


Figure 7: Window analysis of edit-level and sentence-level in meta-evaluation when window size is four. The number in the concentric circumferences represents the Pearson coefficients r, which are increasing from the center of the circle along the radius. Nine numbers x outside the circle represents the consecutive ranking ranges. For example, $x = 4$ involves calculating the Pearson correlation r using the systems ranked from 1st to 4th in the human rankings.

evaluators, the results are coarse-granularity, unexplainable and may have potential subjective bias.

Instability. Based on (Kobayashi et al., 2024a)'s experiment setup, we reproduce the system-level meta-evaluation of GPT-4 focusing on sentence-level fluency. As shown in Table 5 in Appendix, GPT-4 as a scorer highly volatiles because its scoring is relatively subjective, variable and lacking in a standardized scoring rule and method, especially in the SEEDA-E. Therefore, the score of GPT-4 and other LLMs are currently not suitable to be utilized as a metric because of instability and lacking standardized rule, even though they have higher correlation with human judgments in some specific levels.

Position-related biases. Given an original ungrammatical sentence and its five correction sentences, we explore the position-related biases by letting GPT-4 scoring the five corrections only with different permutations in the prompt. Figure 8 in Appendix shows that the score distribution across five corrections and it is obvious that GPT-4's scoring is very instable and has severe position-related biases.

In terms of GEC evaluation metric, our metric has advantages of stability, high-correlation, fined-grained and explainability, which is more suitable for a standardized GEC metric for benchmarking.

5 Benchmark

Most mainstream GEC benchmarks, such as CoNLL14 (Ng et al., 2014), BEA19 (Bryant et al., 2019), and JFLEG (Napoles et al., 2017), rely on conventional metrics like M^2 (Dahlmeier and Ng, 2012), ERRANT (Bryant et al., 2017), and GLEU (Napoles et al., 2015). These metrics exhibit limited correlation with human judgments in meta-evaluation experiments (see Table 1), thereby failing to accurately represent model performance. Although numerous new metrics (Gong et al., 2022; Islam and Magnani, 2021; Maeda et al., 2022; Yoshimura et al., 2020) with higher human correlation have been proposed in recent years, most are only evaluated on meta-evaluation datasets (Grundkiewicz et al., 2015; Kobayashi et al., 2024b) and have not yet been applied to existing GEC benchmarks to produce updated rankings that better align with human evaluations. To address this gap, we apply our proposed metric to the CoNLL14 benchmark—which includes references annotated by two annotators—and evaluate model performance at both the edit level and sentence level.

Table 3 presents the revised system rankings and scores obtained by applying our metric to partial system outputs on the CoNLL14 benchmark. Model performance was originally evaluated using the M^2 scorer on the CoNLL14 leaderboard[5]. As anticipated, the rankings differ signif-

[5] https://paperswithcode.com/sota/grammatical-error-correction-on-conll-2014

Rank	Edit-Level		Sentence-Level		Original	
	System	Score	System	Score	System	Score
1	GRECO(voting+ESC) (⇑2)	75.38	GRECO(voting+ESC) (⇑2)	75.63	Ensembles of best 7 models + GRECO + GTP-rerank	72.8
2	Ensembles of best 7 models + GRECO + GTP-rerank(⇓1)	75.20	Transformer + Pre-train with Pseudo Data (+BERT)(⇑8)	75.41	Majority-voting ensemble on best 7 models	71.8
3	MoECE(⇑4)	75.02	Ensembles of best 7 models + GRECO + GTP-rerank(⇓2)	75.38	GRECO (voting+ESC)	71.12
4	Transformer + Pre-train with Pseudo Data (+BERT)(⇑6)	74.97	Majority-voting ensemble on best 7 models(⇓2)	75.04	GEC-DI (LM+GED)	69.6
5	Majority-voting ensemble on best 7 models(⇓3)	74.84	Transformer + Pre-train with Pseudo Data(⇑6)	75.02	Unsupervised GEC + cLang8	69.9
6	Transformer + Pre-train with Pseudo Data(⇑5)	74.54	SynGEC(⇑2)	74.70	ESC	69.51
7	ESC(⇓1)	74.42	Unsupervised GEC + cLang8(⇓2)	74.66	MoECE	67.79
8	SynGEC(√)	74.01	GEC-DI (LM+GED)(⇓4)	74.64	SynGEC	67.6
9	Unsupervised GEC + cLang8(⇓4)	73.91	Transformer(⇑3)	74.62	GECToR	66.5
10	GEC-DI (LM+GED)(⇓6)	73.90	GECToR(⇓1)	74.58	Transformer + Pre-train with Pseudo Data (+BERT)	65.2
11	GECToR(⇓2)	73.85	MoECE(⇓4)	74.31	Transformer + Pre-train with Pseudo Data	65.0
12	Transformer(√)	73.58	ESC(⇓6)	74.07	Transformer	55.8
13	CNN Seq2Seq(√)	71.18	CNN Seq2Seq(√)	72.74	CNN Seq2Seq	54.79
14	INPUT(√)	59.60	INPUT(√)	64.65	INPUT	0
Δ	**34**		**42**		**-**	

Table 3: System rankings and scores of applying our metric to partial system outputs on CoNLL14 benchmark. **Original** presents the original rankings and M^2 scores on CoNLL14 benchmark leaderborad. (⇑/⇓) denotes that the rank given by the our metric is higher/lower than the original rank, and (√) denotes that the given rank is equal to the original rank. Total rank difference is represented by Δ. According to the results of meta-evaluation and the optimal α and γ determination from Table 2, we select (α=0.195, γ=0.825) for edit-level evaluation and (α=0.035, γ=0.895) for sentence-level such that both of the granularities focus on different evaluation level and reach relatively higher correlations with human judgments. The score processing details are provided in the appendix.

icantly from the original results, both at the edit and sentence levels. However, the total ranking variation is smaller at the edit level, suggesting that the M^2 metric prioritizes edit-level evaluation while partially neglecting sentence-level assessment. Additionally, rankings fluctuate less dramatically for models with either exceptionally strong performance (e.g., the top three models in the original rankings) or notably weak performance (e.g., the bottom three models). In contrast, models with mediocre performance exhibit greater ranking variability. This indicates that while both the conventional M^2 metric and our proposed WODAFE metric can distinguish extreme performance levels, WODAFE aligns more closely with human judgments when differentiating between similar or intermediate-performing models.

The substantial divergence in rankings underscores the limitations of M^2, a conventional metric with relatively low human correlation, in accurately discriminating model performance. Consequently, adopting newer metrics with higher human correlation is critical for establishing more informative and reliable baselines. Our metric, WODAFE—which outperforms most existing metrics in meta-evaluation experiments—provides rankings that better reflect human evaluations, thereby offering a more meaningful and accurate reference for GEC benchmarking.

6 Conclusion

We present WODAFE, a hybrid GEC evaluation metric that addresses the limitations of conventional reference-based and reference-less approaches by jointly optimizing grammaticality, faithfulness, and fluency. By decoupling overcorrection penalties via OCScore and integrating sentence-level fluency assessment, WODAFE achieves superior alignment with human judgments compared to existing metrics. Our analysis highlights two insights: first, discriminative weighting of overcorrection errors significantly reduces biases due to lack of sufficient references, and second, benchmarks like CoNLL14 exhibit evaluation inconsistencies when analyzed through human-centric metrics and conventional metrics. These findings emphasize the need for interpretable, multi-level evaluation frameworks in GEC, particularly as LLMs ineradicable overcorrection tendencies.

7 Limitations

While WODAFE demonstrates strong edit-level performance, its sentence-level fluency score exhibits instability in specific ranking ranges, likely due to the generative nature of reference-less evaluation. The optimal hyperparameters (α, γ) for combining edit- and sentence-level scores may require corpus-specific tuning, introducing practical deployment challenges. Additionally, although WODAFE mitigates biases, inherent limitations of reference-based and reference-less paradigms—such as incomplete references and model hallucination—cannot be fully eliminated.

References

Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. The BEA-2019 shared task on grammatical error correction. In *Proceedings* *of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy. Association for Computational Linguistics.

Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 793–805, Vancouver, Canada. Association for Computational Linguistics.

Christopher Bryant, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe. 2023. Grammatical error correction: A survey of the state of the art. *Computational Linguistics*, 49(3):643–701.

Shamil Chollampatt and Hwee Tou Ng. 2018. A reassessment of reference-based grammatical error correction metrics. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2730–2741, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Leshem Choshen and Omri Abend. 2018a. Automatic metric validation for grammatical error correction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1372–1382, Melbourne, Australia. Association for Computational Linguistics.

Leshem Choshen and Omri Abend. 2018b. Inherent biases in reference-based evaluation for grammatical error correction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 632–642, Melbourne, Australia. Association for Computational Linguistics.

Daniel Dahlmeier and Hwee Tou Ng. 2012. Better evaluation for grammatical error correction. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 568–572, Montréal, Canada. Association for Computational Linguistics.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2022. On the limitations of reference-free evaluations of generated text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10960–10977, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tao Fang, Shu Yang, Kaixin Lan, Derek F. Wong, Jinpeng Hu, Lidia S. Chao, and Yue Zhang. 2023. Is chatgpt a highly fluent grammatical error correction system? a comprehensive evaluation. *Preprint*, arXiv:2304.01746.

Tao Ge, Furu Wei, and Ming Zhou. 2018. Reaching human-level performance in automatic grammatical error correction: An empirical study. *arXiv preprint arXiv:1807.01270*.

Peiyuan Gong, Xuebo Liu, Heyan Huang, and Min Zhang. 2022. Revisiting grammatical error correction evaluation and beyond. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6891–6902, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Takumi Gotou, Ryo Nagata, Masato Mita, and Kazuaki Hanawa. 2020. Taking the correction difficulty into account in grammatical error correction evaluation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2085–2095, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Roman Grundkiewicz, Marcin Junczys-Dowmunt, and Edward Gillian. 2015. Human evaluation of grammatical error correction systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 461–470, Lisbon, Portugal. Association for Computational Linguistics.

Md Asadul Islam and Enrico Magnani. 2021. Is this the end of the gold standard? a straightforward reference-less grammatical error correction metric. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3009–3015, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Masamune Kobayashi, Masato Mita, and Mamoru Komachi. 2024a. Large language models are state-of-the-art evaluator for grammatical error correction. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 68–77, Mexico City, Mexico. Association for Computational Linguistics.

Masamune Kobayashi, Masato Mita, and Mamoru Komachi. 2024b. Revisiting meta-evaluation for grammatical error correction. *Transactions of the Association for Computational Linguistics*, 12:837–855.

Masamune Kobayashi, Masato Mita, and Mamoru Komachi. 2024c. Revisiting meta-evaluation for grammatical error correction. *Preprint*, arXiv:2403.02674.

Mengsay Loem, Masahiro Kaneko, Sho Takase, and Naoaki Okazaki. 2023. Exploring effectiveness of gpt-3 in grammatical error correction: A study on performance and controllability in prompt-based methods. *Preprint*, arXiv:2305.18156.

Koki Maeda, Masahiro Kaneko, and Naoaki Okazaki. 2022. IMPARA: Impact-based metric for GEC using parallel data. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3578–3588, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. 2015. Ground truth for grammatical error correction metrics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational*

Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 588–593, Beijing, China. Association for Computational Linguistics.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2016. There‘s no comparison: Referenceless evaluation metrics in grammatical error correction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2109–2115, Austin, Texas. Association for Computational Linguistics.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2017. JFLEG: A fluency corpus and benchmark for grammatical error correction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 229–234, Valencia, Spain. Association for Computational Linguistics.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14, Baltimore, Maryland. Association for Computational Linguistics.

Chanjun Park, Yeongwook Yang, Chanhee Lee, and Heuiseok Lim. 2020. Comparison of the evaluation metrics for neural grammatical error correction with overcorrection. *IEEE Access*, 8:106264–106272.

Keisuke Sakaguchi, Matt Post, and Benjamin Van Durme. 2014. Efficient elicitation of annotations for human evaluation of machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 1–11, Baltimore, Maryland, USA. Association for Computational Linguistics.

Jinxiang Xie, Yilin Li, Xunjian Yin, and Xiaojun Wan. 2024. Dsgram: Dynamic weighting sub-metrics for grammatical error correction in the era of large language models. *arXiv preprint arXiv:2412.12832*.

Haihui Yang and Xiaojun Quan. 2024. Alirector: Alignment-enhanced Chinese grammatical error corrector. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2531–2546, Bangkok, Thailand. Association for Computational Linguistics.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2021. LM-critic: Language models for unsupervised grammatical error correction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7752–7763, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jingheng Ye, Yinghui Li, Qingyu Zhou, Yangning Li, Shirong Ma, Hai-Tao Zheng, and Ying Shen. 2023. CLEME: Debiasing multi-reference evaluation for grammatical error correction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6174–6189, Singapore. Association for Computational Linguistics.

Jingheng Ye, Zishan Xu, Yinghui Li, Xuxin Cheng, Linlin Song, Qingyu Zhou, Hai-Tao Zheng, Ying Shen, and Xin Su. 2024. Cleme2.0: Towards more interpretable evaluation by disentangling edits for grammatical error correction. *Preprint*, arXiv:2407.00934.

Ryoma Yoshimura, Masahiro Kaneko, Tomoyuki Kajiwara, and Mamoru Komachi. 2020. SOME: Reference-less sub-metrics optimized for manual evaluations of grammatical error correction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6516–6522, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Wei Zhao, Liang Wang, Kewei Shen, Ruoyu Jia, and Jingming Liu. 2019. Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data. 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 156–165, Minneapolis, Minnesota. Association for Computational Linguistics.

A Ablation Experiment of Fluency Metric

	SEEDA-E		**SEEDA-S**	
	r	ρ	r	ρ
M^2-BASE	0.791	0.764	0.658	0.487
M^2-Flu.	-0.239	0.161	-0.336	-0.013
ERRANT-BASE	0.697	0.671	0.557	0.406
ERRANT-Flu.	-0.502	0.051	-0.587	-0.116
GLEU-BASE	0.911	0.897	0.847	**0.886**
GLEU-Flu.	0.053	0.482	-0.039	0.475
fluency-BASE	0.904	0.902	0.760	0.720
fluency-Flu.	**0.962**	**0.938**	**0.923**	0.824
	Acc	τ	*Acc*	τ
M^2-BASE	0.582	0.328	0.512	0.200
M^2-Flu.	0.527	0.216	0.496	0.170
ERRANT-BASE	0.573	0.310	0.498	0.189
ERRANT-Flu.	0.511	0.188	0.471	0.129
GLEU-BASE	0.695	0.404	0.673	0.351
GLEU-Flu.	0.630	0.266	0.611	0.227
fluency-BASE	**0.755**	**0.511**	**0.747**	**0.493**
fluency-Flu.	0.752	0.505	0.728	0.457

Table 4: Correlation coefficients between human judgments and fluency metric. Upper part of the table is system-level meta-evaluation, while the lower part of the table is sentence-level meta-evaluation. **Bold** indicates the maximum coefficients across each column.

B Instability and Position-related biases of GPT-4 Scoring

	SEEDA-E		SEEDA-S	
	r	ρ	r	ρ
GPT-4-BASE*	**0.974**	0.979	0.913	0.874
GPT-4-BASE′	0.948	0.902	**0.923**	0.916
GPT-4-BASE″	0.944	0.874	0.918	**0.923**
Ours-BASE	0.931	**1.0**	0.886	0.895
GPT-4-Flu.*	**0.981**	0.982	**0.952**	0.916
GPT-4-Flu.′	0.962	0.934	0.937	0.943
GPT-4-Flu.″	0.965	0.916	0.941	**0.947**
Ours-Flu.	0.962	**0.987**	0.923	0.895

Table 5: Results of system-level meta-evaluation of reproducing GPT-4 as a scorer focusing on sentence-level fluency. * represents the results from the original paper, ′ and ″ are two results we reproduced.

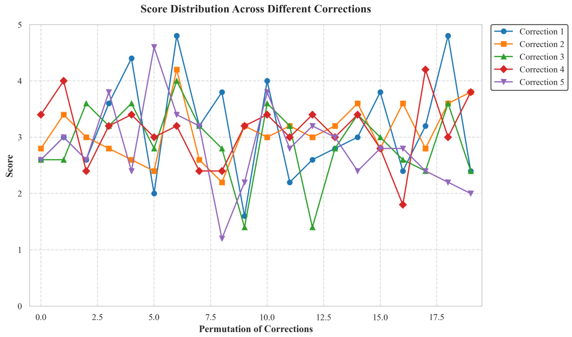


Figure 8: gpt-4-1106-preview score distribution across five different corrections. Horizontal axis represents 20 different permutations of the five corrections in the prompt. Our prompt and experiment setting is in align with previous works's protocol (Kobayashi et al., 2024a).

C Final Score Processing

Edit-level generalized F-score $F_{\beta-G}$ is directly used as one part of final score. Sentence-level fluency, however, has a problem that the scores are relatively centralized and with less differentiation as Figure 9 shows. Given a fair number of scores are centralized in the average score of 0.19 and do not surpass 0.25, we multiply the original fluency scores by four to widen the difference between scores and use the score with multiplication as one part of the final score. If the fluency score of a sentence is greater than 1, we normalize this score to 1.0.

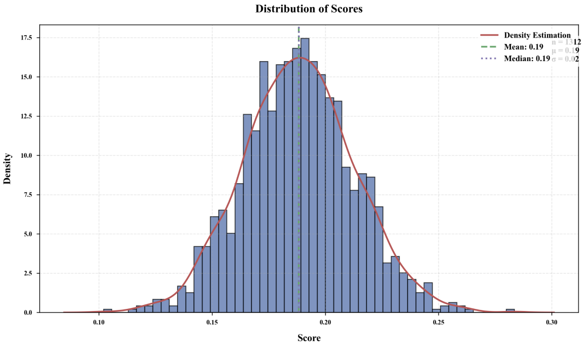


Figure 9: Distribution of the original fluency scores in a model output of CoNLL14 test set.

D Related Work

D.1 GEC Metrics

A core component of a GEC system is the ability to measure model performance. **Reference-based metrics**, the traditional paradigm, rely on alignment with human-crafted references. Early approaches like the M^2 scorer used edit-based $F_{0.5}$ scoring but faced criticism for artificially inflated precision (Bryant et al., 2017), prompting linguistically grounded alternatives like *ERRANT* (Bryant et al., 2017) for improved edit alignment. Despite their effectiveness, reference-based metrics often suffer from issues such as overfitting to reference-specific patterns and inevitably penalizing valid but non-reference corrections, which introduces inherent biases (Choshen and Abend, 2018b). To circumvent reliance on references, **reference-less metrics** emerged, leveraging quality estimation frameworks: Grammaticality-Based Metrics (Napoles et al., 2016) combined fluency and meaning preservation scores, later extended by SOME (Yoshimura et al., 2020) and IMPARA (Maeda et al., 2022) with BERT-based models and Scribendi Score (Islam and Magnani, 2021) using perplexity-edit hybrids. However, these metrics, while bypassing the need for references, can lack interpretability and may inherit biases from the models they employ (Deutsch et al., 2022), and struggle to detect issues like over-correction. **The advent of LLMs** introduced a paradigm shift, with GPT-4 (Kobayashi et al., 2024a) and Prompt Engineering (Xie et al., 2024) enabling direct numerical scoring via natural language criteria interpretation, aiming to mimic human-like evaluation and potentially capturing more complex aspects of correction quality. Despite their promise, LLM-based metrics face challenges in explainability (acting as "black-box" evaluators), instability, subjective bias, and prohibitive computational costs. Ultimately, **meta-evaluation** frameworks determine metric validity: human correlation. This involves evaluating metrics themselves by correlating their scores with human judgments based on two dominant datasets GJG15 (Grundkiewicz et al., 2015) and SEEDA (Kobayashi et al., 2024b), providing a gold standard for determining the true effectiveness of GEC evaluation methodologies and guiding the development of superior metrics.

D.2 Overcorretion

(Fang et al., 2023; Loem et al., 2023) have found that there is a large amount of overcorrection in LLMs. One of the main reasons of overcorrection is that models tend to generate sentences with higher probabilities and replace infrequent words with more frequent ones (Yang and Quan, 2024). Previous works (Yang and Quan, 2024; Zhao et al., 2019) explore various approaches to relieve this problem. In terms of metrics for evaluating overcorrection, (Park et al., 2020) uses levensteins algorithm and the longest common substring (LCS) algorithm, while (Ye et al., 2024) measures overcorrection by disentangling FP and calculating specific ratio.

E Heat Maps

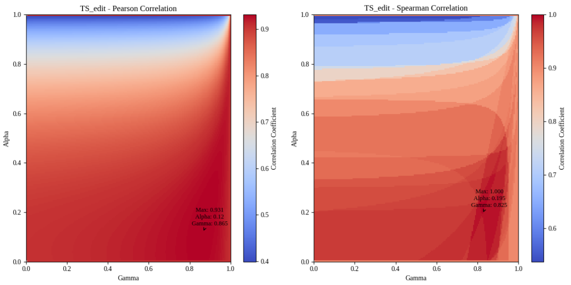


Figure 10: Heat map of system-level SEEDA-E. (Two Annotators)

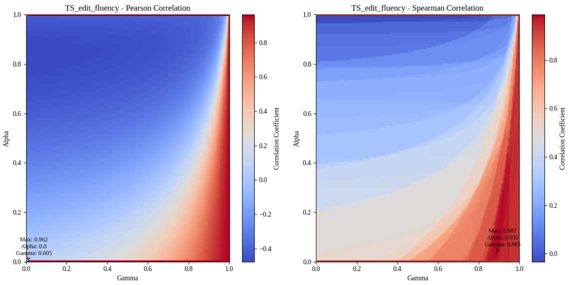


Figure 11: Heat map of system-level SEEDA-E. (+Fluent Corr.) (Two Annotators)

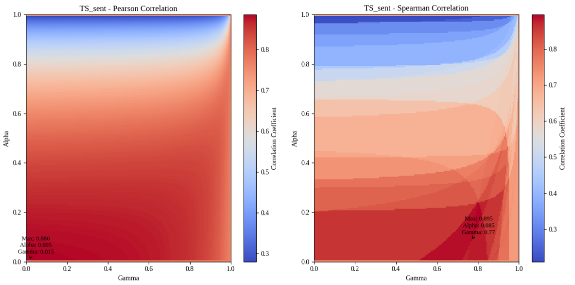


Figure 12: Heat map of system-level SEEDA-S. (Two Annotators)

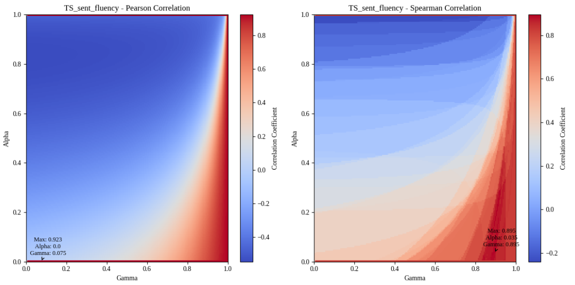


Figure 13: Heat map of system-level SEEDA-S. (+Fluent Corr.) (Two Annotators)

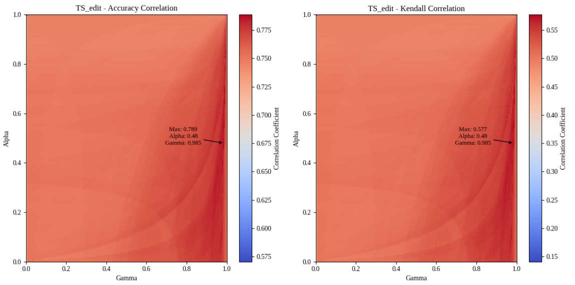


Figure 14: Heat map of sentence-level SEEDA-E. (Two Annotators)

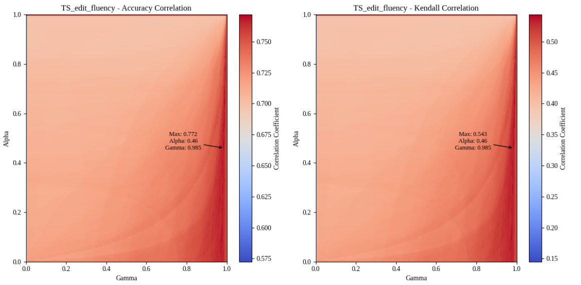


Figure 15: Heat map of sentence-level SEEDA-E. (+Fluent Corr.) (Two Annotators)

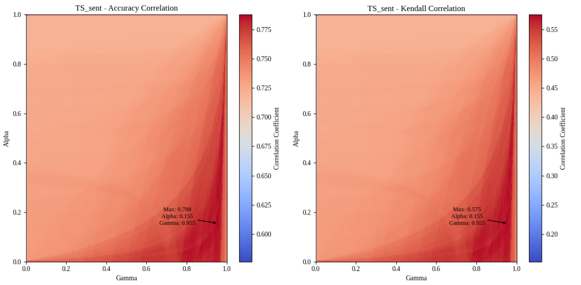


Figure 16: Heat map of sentence-level SEEDA-S. (Two Annotators)

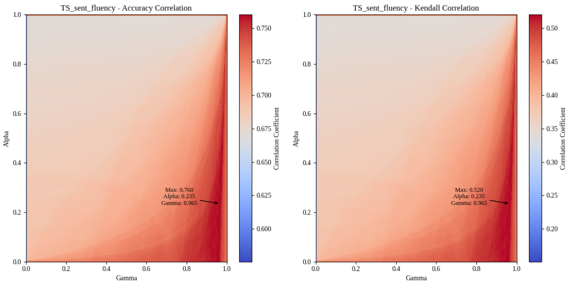


Figure 17: Heat map of sentence-level SEEDA-S. (+Fluent Corr.) (Two Annotators)

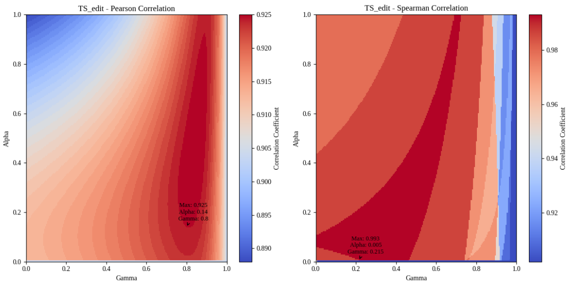


Figure 18: Heat map of system-level SEEDA-E. (Ten Annotators)

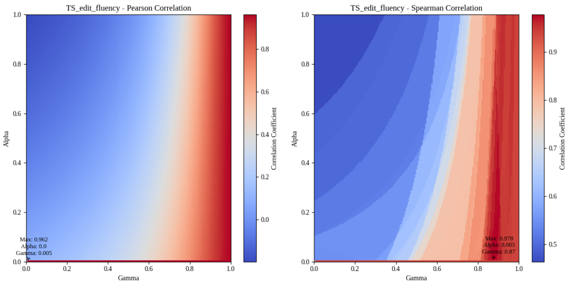


Figure 19: Heat map of system-level SEEDA-E. (+Fluent Corr.) (Ten Annotators)

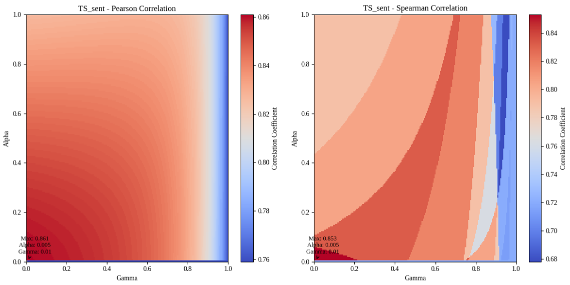


Figure 20: Heat map of system-level SEEDA-S. (Ten Annotators)

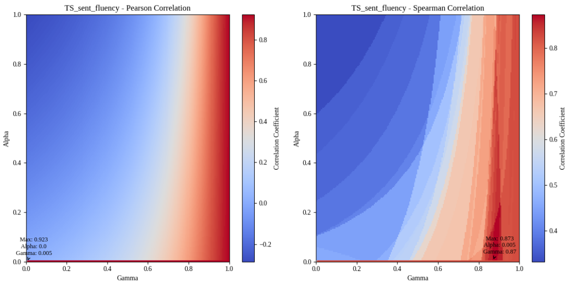


Figure 21: Heat map of system-level SEEDA-S. (+Fluent Corr.) (Ten Annotators)

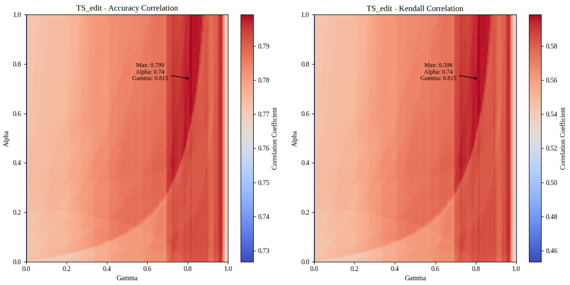


Figure 22: Heat map of sentence-level SEEDA-E. (Ten Annotators)

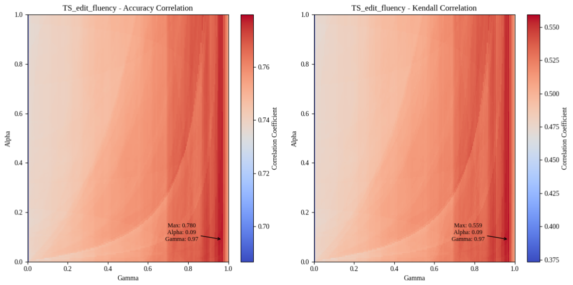


Figure 23: Heat map of sentence-level SEEDA-E. (+Fluent Corr.) (Ten Annotators)

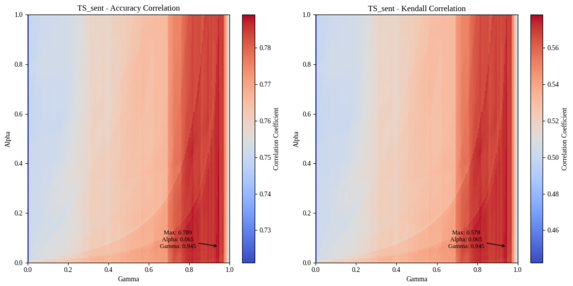


Figure 24: Heat map of sentence-level SEEDA-S. (Ten Annotators)

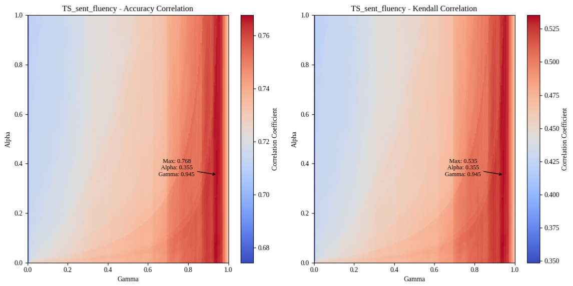


Figure 25: Heat map of sentence-level SEEDA-S. (+Fluent Corr.) (Ten Annotators)