DSGram: Dynamic Weighting Sub-Metrics for Grammatical Error Correction in the Era of Large Language Models

Anonymous submission

Abstract

Evaluating the performance of Grammatical Error Correction (GEC) models has become increasingly challenging, as large language model (LLM)-based GEC systems often produce corrections that diverge from provided gold references. This discrepancy undermines the reliability of traditional reference-based evaluation metrics. In this study, we propose a novel evaluation framework for GEC models, DSGram, integrating Semantic Coherence, Edit Level, and Fluency, and utilizing a dynamic weighting mechanism. Our framework employs the Analytic Hierarchy Process (AHP) in conjunction with large language models to ascertain the relative importance of various evaluation criteria. Additionally, we develop a dataset incorporating human annotations and LLM-simulated sentences to validate our algorithms and fine-tune more cost-effective models. Experimental results indicate that our proposed approach enhances the effectiveness of GEC model evaluations.

Introduction

Grammatical Error Correction (GEC) models aims to automatically correct grammatical errors in natural language texts, enhancing the quality and accuracy of written content. Traditionally, the evaluation of GEC models has employed a variety of metrics, categorized into those requiring a reference (reference-based evaluation) and those that do not (reference-free evaluation).

Reference-based metrics such as BLEU (Papineni et al. 2002), ERRANT (Bryant, Felice, and Briscoe 2017), and M^2 (Dahlmeier and Ng 2012) compare the model-generated text with a correct reference text to evaluate the accuracy of grammatical corrections, and they are widely used in this area. Despite the usefulness of these metrics, they possess inherent limitations. For example, the golden reference may not encompass all potential corrections (Choshen and Abend 2018), and the alignment of existing automatic evaluation metrics with human judgment is often weak (Coyne et al. 2023). Additionally, LLMs-based GEC models may excessively correct sentences, resulting in unnecessary editing not captured by traditional metrics (Fang et al. 2023b).

Conversely, reference-free metrics like Perplexity, GLEU (Napoles et al. 2015), and SOME (Yoshimura et al. 2020) assess the quality of generated text directly, proving beneficial when a correct reference text is unavailable, or when a single (or several) reference is insufficient for GEC evaluation in the era of LLMs. While existing reference-free evaluation metrics like SOME offer an analytical foundation, they no longer fully encompass the scope of GEC evaluation. It is necessary to design new sub-metrics based on the original sub-metrics and emerging requirements. Figure 1 shows running examples of representative GEC systems and the limitations of current evaluation metrics.

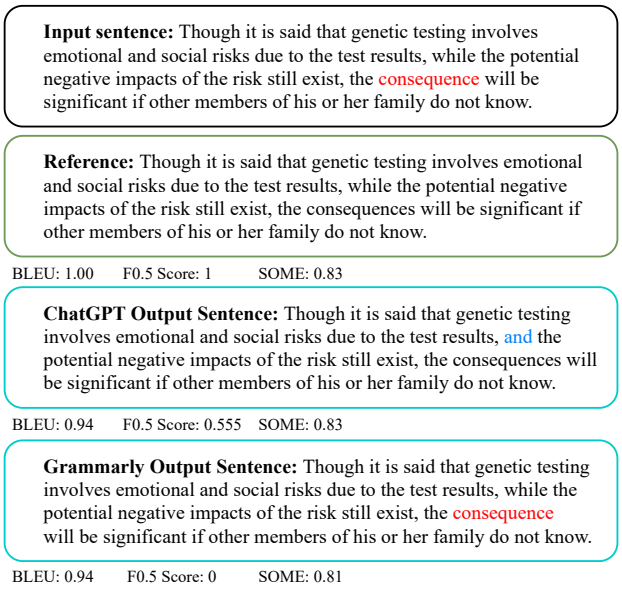

Figure 1: Running examples and evaluation results of several existing metrics. The sentence is from Fang et al. (2023b). This figure illustration presents the outcomes of an input sentence processed through two representative GEC models and also the corresponding reference. The metrics' scores are placed under the output. The highlighted in blue represents the over-correction. The highlighted in red indicates poor fluency. Notably, BLEU fails to differentiate between over- and under-correction, whereas SOME cannot capture over-correction.

This study seeks to overcome these limitations by introducing a novel evaluation framework, DSGram, for GEC models. DSGram integrates Semantic Coherence, Edit Level, and Fluency, and determines the appropriate weights

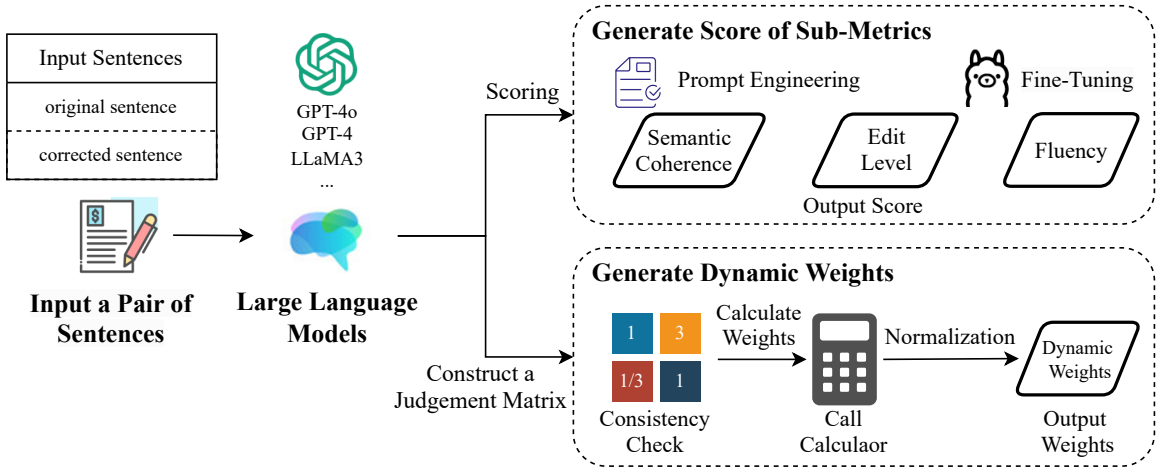

Figure 2: Architecture of the DSGram method. It begins with the input of sentence pairs (original and corrected). Large language models (such as GPT-4 and LLaMA3) are employed to generate dynamic weights, which are further refined through a judgment matrix and a consistency check. These dynamic weights are normalized and then used to score sub-metrics, which include semantic coherence, edit level, and fluency. The final output consists of both the calculated weights and the overall score for the evaluated correction.

for these criteria. By initially training a weighting system using human-annotated datasets and later adjusting these weights based on the evaluation scenario's context, a more nuanced and context-sensitive evaluation approach can be developed.

The main contributions of this paper are summarized as follows:

- We introduce new sub-metrics for GEC evaluation, diverging from previous categorical approaches. Our metrics optimize past sub-metrics by adding an evaluation of over-editing.
- We propose a novel dynamic weighting-based GEC evaluation method, DSGram, which integrates the Analytic Hierarchy Process (Ana 1987) with large language models to ascertain the relative importance of different evaluation criteria.
- We present two datasets: DSGram-Eval, created through human scoring, and DSGram-LLMs, a larger dataset designed to simulate human scoring for fine-tuning. Both datasets utilize sentences from the CoNLL-2014 and BEA-2019 test sets to facilitate rigorous evaluation.[1]

Related Work

Numerous studies have focused on evaluating GEC models. This section provides an overview of key research utilizing LLMs for GEC assessment.

Model-Based Evaluation Metrics

Model-based evaluation metrics have garnered significant attention, especially in the domain of GEC. BLEURT (Sellam, Das, and Parikh 2020) is a versatile metric that evaluates based on (prediction, reference) pairs. It employs a scalable pretraining phase where it learns to predict automatically generated signals of supervision from semantically comparable synthetic pairs.

A burgeoning trend in automatic evaluation involves the direct application of LLMs for assessment purposes. Liu et al. (2023) have LLMs generate numerical ratings by interpreting descriptions of evaluation criteria through a chain-of-thought method (Wei et al. 2023). Sottana et al. (2023) demonstrate the viability of GPT-4 in GEC assessment, highlighting an approach that uses natural language instructions to define evaluation criteria.

Reference-Free Evaluation

Reference-free evaluation is a method of assessing the performance of models without relying on reference. Asano, Mizumoto, and Inui (2017) integrate three sub-metrics, which are Grammaticality, Fluency, and Meaning Preservation, to surpass reference-based metrics. They employ a language model and edit-level metrics as Fluency and Meaning Preservation sub-metrics, respectively, although these sub-metrics are not tailored for manual evaluation. The final score is determined through a weighted linear summation of each individual evaluation score.

SOME is a reference-free GEC metric that follows the approach of Asano, Mizumoto, and Inui (2017). It utilizes three distinct BERT models, each dedicates to one scoring aspect. The researchers construct a novel dataset for training these BERT models by annotating the outputs from various GEC systems on the CoNLL-2013 test set across the three scoring dimensions.

However, in real-world applications, the use of just three evaluation metrics presents a limitation, as it hinders the

[1]Our code, models and data will be released to the community to facilitate future research.

generation of a holistic score in an inherently intuitive way. Consequently, this constraint diminishes its effectiveness in directing the training of GEC models. In the sphere of GEC evaluation, Wu et al. (2023) assessed ChatGPT and observed a tendency towards over-correction, which might be attributed to its extensive generative capacities as a large language model. Similarly, Fang et al. (2023b) noted that ChatGPT makes over-corrections, leading to revised sentences with high fluency. These findings align with the outcomes of our own experiments, which indicate that existing reference-free metrics do not adequately reflect these issues.

GEC Dataset with Human Scoring

There are very few GEC datasets with human evaluation scores. The dataset annotated by Yoshimura et al. (2020) includes individual scores for Grammaticality, Fluency, and Meaning Preservation, but lacks an overall score. GJG15 (Grundkiewicz, Junczys-Dowmunt, and Gillian 2015) and SEEDA (Kobayashi, Mita, and Komachi 2024) are manually annotated but provide ranking information rather than scores. Sottana et al. (2023) conducted ratings based on the gold standard, which presents certain limitations.

Sottana et al. (2023) suggested the potential of GPT-4 as a reviewer, prompting us to consider using GPT-4 to annotate a dataset that simulates human scoring for GEC model evaluation.

DSGram: Dynamic Weighting Sub-Metrics for GEC

DSGram comprises two main components: score generation and weight generation. By applying specific weights to the generated scores, an overall score is obtained. Figure 2 illustrates the method's flowchart.

Generating Scores

Sub-Metrics Definition Upon our analysis, the three sub-metrics introduced by Asano, Mizumoto, and Inui (2017), exhibit redundancy and insufficiency. We compute the correlation of these metrics using the SOME's dataset, and the results are depicted in Figure 3.

The heatmap reveals a high correlation between Grammaticality and Fluency. Hence, we have combined these two metrics into a single "Fluency" measure. Furthermore, to address the issue of over-corrections present in LLM-based GEC models, we have incorporated a new sub-metric called "Edit Level" to evaluate concerns associated with excessive corrections. Based on the computation of 200 human-annotated scores, The correlation of these novel sub-metrics is shown in Figure 4. It can be observed that our classification criteria have improved the sub-metrics' distribution.

In our definition, the meanings of the three sub-metrics are as follows:

Semantic Coherence The degree to which the meaning of the original sentence is preserved in the corrected sentence. It evaluates whether the corrected sentence conveys the same intended meaning as the original, without introducing semantic errors or altering the core message.

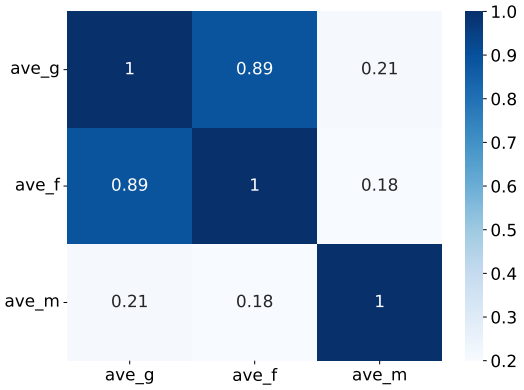

Figure 3: Heat map of the correlations among the three sub-metrics of SOME. avg_g, avg_f, and avg_m denote the average scores for Grammaticality, Fluency, and Meaning Preservation, respectively. The correlation between avg_g and avg_f is notably strong at 0.89.

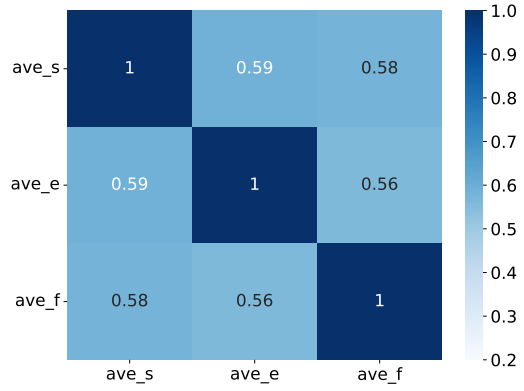

Figure 4: Heat map of our three sub-metrics correlations. avg_s, avg_e, and avg_f denote the average scores for Semantic Coherence, Edit Level and Fluency, respectively. The correlation has become more evenly distributed.

Edit Level The extent to which the GEC model has modified the sentence. It assesses whether the corrections made are necessary and appropriate, or if the sentence has been unnecessarily or excessively altered, deviating from the original prose more than required.

Fluency The grammatical correctness and the natural flow of the corrected sentence. It evaluates whether the sentence adheres to proper grammar rules, has a coherent structure, and reads smoothly without awkward phrasing or unnatural constructions.

These three sub-metrics comprehensively cover the key aspects that GEC models need to consider in the era of large language models.

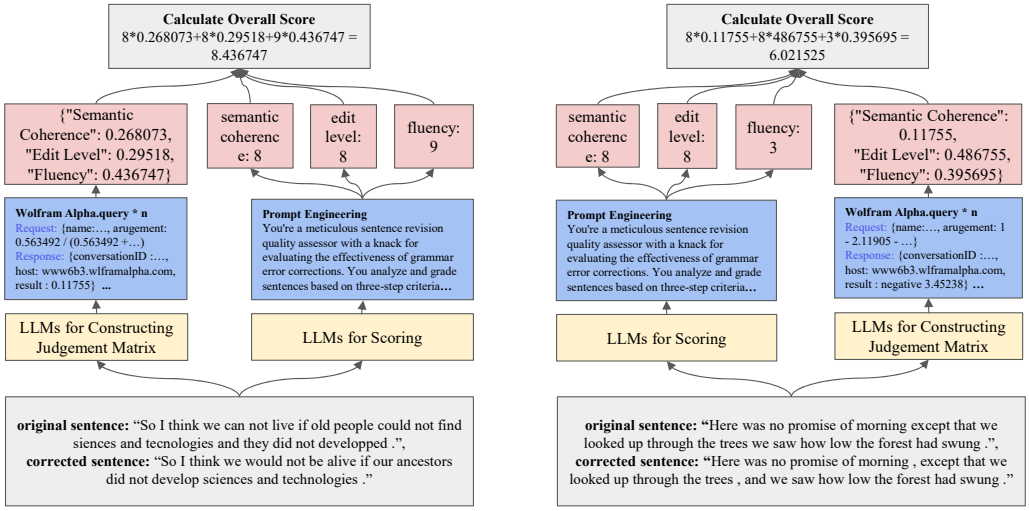

Figure 5: The DSGram score computation processes for two different sentences. Sentence (a) represents a casual daily dialogue where the emphasis is on Fluency. Sentence (b) is a more formal expression, thus placing a greater emphasis on the Edit Level.

Sub-Metrics Score Generation We employ prompt engineering techniques such as chain-of-thought (Wei et al. 2023), few-shot (Brown et al. 2020) and output reason before scoring (Chu, Chen, and Nakayama 2024) to craft the prompts, enabling LLMs to generate scores based on the aforementioned three sub-metrics. The specific prompts can be found in Appendix.

The method is applicable to a variety of different LLMs. Our experiments reveal that the GPT-4 and GPT-4o exhibit greater scoring consistency and alignment with human judgments. Models such as LLaMA2 and LLaMA3 struggle to adhere to structured prompts and do not effectively grasp the scoring tasks. To reduce evaluation costs and enhance model usability, we annotate a dataset simulating human scoring with GPT-4 to fine-tune open-source LLMs, exploring their feasibility for scoring. We utilize GPT-4 to annotate DSGram-LLMs and subsequently fine-tune the LLaMA2-13B and LLaMA3-8B models on this dataset, thereby confirming their consistency with human scores.

In summary, we achieve the automatic generation of evaluation scores with prompting GPT-4 and open-source LLMs like LLaMA.

Generating Dynamic Weights

The prevalent approach using sub-metrics methods typically involves assigning specific weights to compute an overall score. However, we contend that in the context of GEC evaluation tasks, human judges may have varying considerations for different sentences. It is essential to adjust the weights according to the evaluation scenario; for formal documents such as legal texts and medical instructions, a stricter standard is required for Semantic Coherence and Edit Level. In contrast, for more relaxed contexts like dialogues, Fluency should be given priority in the assessment.

We employ LLMs in conjunction with the Analytic Hierarchy Process (AHP) (Ana 1987) to generate varying weights for different sentences, thereby making the weighted scores more aligned with human scoring. AHP is a decision-making framework that decomposes complex decisions into a hierarchy of simpler, more manageable elements through pairwise comparisons. It synthesizes subjective judgments into objective priorities that are appropriately weighted based on a structured evaluation of criteria and objectives.

The weight generation algorithm involves the following key steps:

Constructing the Judgement Matrix: For each criterion, utilize LLMs to construct a pairwise comparison matrix $A = (a_{ij})$, where a_{ij} represents the importance of criterion i relative to criterion j. The specific prompt for this process is documented in the Appendix.

Consistency Check: Compute the consistency index (CI) and consistency ratio (CR) to check the consistency of the pairwise comparison matrix.

$$CI = \frac{\lambda_{\max} - n}{n-1},$$

$$CR = \frac{CI}{RI},$$

where RI is the Random Index for a matrix of order n. For instance, an RI value of 0.58 is used for a 3x3 matrix, while for a 4x4 matrix, RI is 0.90, and this pattern continues. A CR below 0.1 indicates that the pairwise comparison matrix possesses sufficient consistency.

Calculating the Eigenvector and Eigenvalue: Solve the eigenvalue problem $A\mathbf{w} = \lambda_{\max}\mathbf{w}$ to obtain the maximum eigenvalue $\lambda_{\max}$ and the corresponding eigenvector $\mathbf{w}$, which is then normalized.

Calculating the Composite Weight: Compute the composite weights by multiplying the weights at each level.

Algorithm 1: Adaptive Weight Calculation for GEC Evaluation using AHP and LLMs

Initialize: Define criteria set $C = C_1, C_2, \ldots, C_n$ (e.g., Semantic Coherence, Edit Level, Fluency)
Construct Judgment Matrices:
each pair of criteria (C_i, C_j), each sentence S_k Construct judgment matrix $A = [a_{ij}]$ utilizing LLMs for S_k
where a_{ij} represents the relative importance of C_i to C_j in the context of this sentence
Consistency Check:
each judgment matrix A Calculate Consistency Index (CI) and Consistency Ratio (CR)
CR $\geq \theta$ Adjust ratings in A until CR $< \theta$
Calculate Weights:
each judgment matrix A Compute the principal eigenvector $\mathbf{w} = (w_1, w_2, \ldots, w_n)$ of A
Normalize the principal eigenvector to obtain weights w_i for each criterion C_i:
Compute Aggregated Scores:
each sentence S_k Calculate weighted score $W_k = \sum_{i=1}^{n} w_i \cdot s_{k,i}$
where $s_{k,i}$ is the score of S_k on criterion C_i
Output: Aggregated scores for each S_k

The procedure for this approach is outlined in Algorithm 1. Figure 5 depicts an illustration of this method for two distinct sentence pairs. Sentence (a) represents a casual daily dialogue where the emphasis is on Fluency. On the other hand, sentence (b) is a more formal expression with a higher level of professionalism, thus placing a greater emphasis on the Edit Level.

Experiments and Analysis

Datasets Preparation

We mainly evaluate our metrics using the SEEDA dataset (Kobayashi, Mita, and Komachi 2024), and compare them with existing metrics. The SEEDA corpus comprises human-rated annotations at two distinct evaluation levels: edit-based (SEEDA-E) and sentence-based (SEEDA-S), and encompasses corrections from 12 cutting-edge neural systems, including large language models, as well as corrections made by humans.

In addition to SEEDA, we also manually annotate an evaluation dataset with human scores for validation purposes. We name this human-annotated dataset DSGram-Eval. For details on annotating this dataset, please refer to the Appendix.

Furthermore, we construct an additional training set by utilizing system outputs from the BEA-2019 shared task official website[2]. After removing biased output groups, a sample of approximately 2500 entries is randomly selected. Then GPT-4 is employed to generate the sub-metrics scores, which we refer to as DSGram-LLMs. This set is subsequently used for model training.

Meta-Evaluation of Our Metrics

We evaluate the performance of various GEC models using the DSGram metric on the SEEDA dataset. GPT-3.5 secures the highest score in our DSGram metric, aligning with the findings in Fang et al. (2023b), which report ChatGPT as having the highest human score in fluency. Despite it exhibiting substandard scores in metrics like M^2, we still regard it as an outstanding GEC system. Our metric demonstrates equally outstanding performance on models like T5 (Rothe et al. 2021) and TransGEC (Fang et al. 2023a) that excel in GLEU and ERRANT. However, the REF-F model, despite boasting the highest SOME score, underperforms significantly when assessed using our metric. The results are shown in Table 1.

We calculate the correlation between the ranking according to DSGram and human ranking on the SEEDA dataset and compare it with existing metrics. Specifically, we convert the concrete data obtained in Table 1 into human rankings and compare them with the human rankings in the dataset, obtaining the corresponding correlation. The comparison is presented in Table 2 and the obtained correlation is presented in Table 3.

The results indicate that the correlation of our metric with human feedback surpasses that of all conventional reference-based metrics, as well as reference-free metrics like GLEU, Scribendi Score and DSGram weighted by 0.33.

Validation of LLMs Scoring

To verify the broad applicability of our metrics, we conduct tests on DSGram-Eval dataset by using various LLMs (including few-shot and finetuned LLaMA models) to generate the sub-metric scores and then calculate the correlation between the sub-metric/overall scores with human scores. The correlation results in Table 4 indicate that GPT-4 consistently shows a high correlation with human scores across three sub-metrics. In contrast, LLaMA3-70B is found to be highly correlated in Semantic Coherence and Fluency but less so in Edit Level. Since the few-shot LLaMA performs poorly, we fine-tune LLaMA on the DSGram-LLMs dataset. The fine-tuned LLaMA3-8B and LLaMA2-13B models outperform their original few-shot outcomes, demonstrating that LLMs can achieve higher human-aligned scores for GEC tasks, while fine-tuning smaller LLMs serves to reduce evaluation costs and enhance model usability.

Validation of Dynamic Weights

To ascertain the efficacy of the dynamic weights in DSGram, we conduct a comparison between the average overall score assigned by human annotators and the score weighted

[2]https://www.cl.cam.ac.uk/research/nl/bea2019st/

GEC Model	M^2	$ERRANT$	$GLEU$	$SOME$	$DSGram$
BERT-fuse (Kiyono et al. 2019)	62.77	58.99	68.5	0.8151	9.3853
GECToR-BERT (Omelianchuk et al. 2020)	61.83	58.05	66.56	0.8016	9.1473
GPT-3.5	53.5	44.12	65.93	0.8379	**9.6310**
PIE (Awasthi et al. 2019)	59.93	55.89	67.83	0.8066	9.0342
REF-F (most fluency references by experts)	47.48	33.24	60.34	**0.8463**	9.0534
REF-M (minimal edit references by experts)	60.12	54.77	67.27	0.8155	9.4661
Riken-Tohoku (Kaneko et al. 2020)	64.74	61.88	68.37	0.8123	9.4442
T5 (Fang et al. 2023a)	65.07	60.65	**68.81**	0.8202	9.4983
TransGEC (Fang et al. 2023a)	**68.08**	**64.43**	70.20	0.8200	9.5711

Table 1: Scores of common GEC models w.r.t various existing metrics and DSGram on SEEDA dataset. The bold part indicates the model with the highest score under each metric. Certain GEC models with low scores have been omitted from this table.

GEC Models	DSGram Score	GEC Models	Human Score
GPT-3.5	9.631	REF-F	0.992
TransGEC	9.571	GPT-3.5	0.743
T5	9.498	T5	0.179
REF-M	9.466	TransGEC	0.175
Riken-Tohoku	9.444	REF-M	0.067
UEDIN-MS	9.411	BERT-fuse	0.023
BERT-fuse	9.385	Riken-Tohoku	-0.001
GECToR-BERT	9.147	PIE	-0.034
REF-F	9.055	LM-Critic	-0.163
PIE	9.034	TemplateGEC	-0.168
GECToR-ens	9.030	GECToR-BERT	-0.178
LM-Critic	9.017	UEDIN-MS	-0.179
BART	8.934	GECToR-ens	-0.234
TemplateGEC	8.899	BART	-0.300
INPUT	8.127	INPUT	-0.992

Table 2: A comparative analysis of the DSGram Score and the Human Score in SEEDA, as ranked by the DSGram Score, indicates a favorable correlation, with our scores exhibiting a close alignment to those assigned by humans.

by generated dynamic weights derived from human-labeled sub-metrics.

We calculate the AHP Human Score by applying dynamic weights to the human-annotated Semantic Coherence, Edit Level, and Fluency scores. The Pearson correlation coefficient between the AHP Human Score and the human-annotated overall score is **0.8764**. Subsequently, we compute the overall score derived from the average weighting method and correlated it with the human scoring overall score. By assigning an equal weight of 0.33 to each metric, we obtain the AVG Human Score. The calculation reveal a Pearson correlation coefficient of **0.8544** between the AVG Human Score and the human scoring overall score.

Our experimental findings indicate that when the overall score is adjusted using dynamic weights, it significantly corresponds with the evaluations provided by human annotators, and the dynamic weighting approach outperforms the average weighting method at a relaxed significance level.

It is suggested that, although human annotators do not consciously consider the specific weights of each metric while annotating, they implicitly gauge their relative importance. This intuitive evaluation process resembles the creation of a judgment matrix, and our method effectively mimics the human scoring procedure.

To further verify the effectiveness, we select three distinctly different text datasets: OpenSubtitles[3], Justia (CaseLaw)[4], and Wikipedia Corpus, to represent everyday conversations, legal documents, and technical explanations, respectively. For each dataset, we generate weights for three sub-metrics and evaluate the internal consistency of these weights using Cronbach's Alpha coefficient (Cronbach 1951). Specifically, we preprocess each dataset to extract sub-metric scores and then calculate Cronbach's Alpha based on the generated weights to assess the consistency and reliability of the sub-metrics across different datasets.

The experimental results demonstrate that the Cronbach's Alpha coefficients for the OpenSubtitles, Justia (CaseLaw), and Wikipedia Corpus datasets are **0.76**, **0.82**, and **0.79**, respectively, all exceeding the acceptable threshold of 0.7.

This indicates that the weights generated in different scenarios exhibit high consistency and reliability. The weights in the everyday conversation scenario show good consistency, while the weights in the legal and technical docu-

[3]https://www.opensubtitles.org/

[4]https://law.justia.com/cases/

Metrics	System-level				Sentence-level			
	SEEDA-S		SEEDA-E		SEEDA-S		SEEDA-E	
	r	ρ	r	ρ	Acc	τ	Acc	τ
M^2	0.658	0.487	0.791	0.764	0.512	0.200	0.582	0.328
$ERRANT$	0.557	0.406	0.697	0.671	0.498	0.189	0.573	0.310
$GLEU$	0.847	0.886	0.911	0.897	0.673	0.351	0.695	0.404
$SOME$	0.892	0.867	0.901	**0.951**	0.768	**0.555**	0.747	**0.512**
$IMPARA$	**0.911**	0.874	0.889	0.944	0.761	0.540	0.742	0.502
Avg_DSGram	0.797	0.790	0.922	0.930	0.648	0.296	0.661	0.323
$DSGram$	0.880	**0.909**	**0.927**	0.944	**0.776**	0.551	**0.750**	0.499

Table 3: System-level and sentence-level meta-evaluation results of common GEC models. We follow the Kobayashi, Mita, and Komachi (2024),use Pearson (r) and Spearman (ρ) for system-level and Accuracy (Acc) and Kendall (τ) for sentence-level meta-evaluations. The sentence-based human evaluation dataset is denoted SEEDA-S and the edit-based one is denoted SEEDA-E. The score in bold represents the metrics with the highest correlation at each granularity.

LLMs	Semantic Coherence	Edit Level	Fluency	Overall
GPT-4 Turbo	**0.724**	**0.839**	**0.797**	**0.772**
LLaMA3-70B	0.399	0.349	0.574	0.607
LLaMA2-13B (Fine-tuned)	0.315	0.239	0.258	0.382
LLaMA3-8B (Fine-tuned)	0.372	0.331	0.361	0.419
LLaMA2-13B (5-shot)	0.215	0.189	0.245	0.306
LLaMA3-8B (5-shot)	0.327	0.243	0.261	0.373

Table 4: Pearson correlation with human scoring on DSGram-Eval across various LLMs

mentation scenarios perform even better. Overall, this experiment validates the effectiveness of the weight generation algorithm across different text types, demonstrating its robustness and reliability in diverse applications.

Discussion

Human Feedback For reference-free evaluation metrics, their effectiveness is often gauged by how closely they mirror human judgments. Although ChatGPT is rated as the best model in terms of grammar by human reviewers, it is rated as the worst or second-worst model in terms of semantics and over-correction (Sottana et al. 2023). This suggests that human feedback also seems to have its limitations.

LLMs as Reviewers The utilization of large language models to construct evaluation metrics has been found to closely approximate human feedback, thereby enhancing the accuracy of assessment. Additionally, the dynamic evaluation approach demonstrates a degree of rationality, which can potentially be transferred to other forms of assessment.

Applications The metrics in question may prove applicable in guiding the training of GEC models. For large-scale models, the rate of gradient descent during the training process may vary across different tasks. For instance, the model's ability to correct semantic errors may have already saturated, whereas there is still room for improvement in terms of fluency correction. Employing this metric to construct the loss function could potentially enhance the training of GEC models.

Conclusions

This study presents an evaluation framework for Grammatical Error Correction models that integrates Semantic Coherence, Edit Level, and Fluency through a dynamic weighting system. By leveraging the AHP in conjunction with LLMs, we have developed a method that dynamically adjusts the importance of different evaluation criteria based on the context, leading to a more nuanced and accurate assessment.

Through extensive experiments, our methodology has demonstrated strong alignment with human judgments, particularly when utilizing advanced LLMs like GPT-4. The dynamic weighting system has shown considerable promise in mirroring the intuitive scoring processes of human annotators, thereby validating its application in various contexts.

Looking forward, several avenues for future research are evident. First, it is imperative to conduct more extensive tests of this method across a broader range of LLMs, including Claude, GLM and Qwen. Moreover, a more comprehensive validation of the dynamic evaluation approach is required, to explore its applicability in diverse evaluation contexts.

However, GPT-4, utilized in creating judgment matrices, is not freely available to the public and might require specific authorization or payment for its use. This limitation could hinder the broad adoption of our evaluation method. Moreover, the performance of the fine-tuned LLaMA model is suboptimal and might necessitate additional high-quality data for effective fine-tuning.

References

1987. The Analytic Hierarchy Process—What It Is and How It Is Used. *Mathematical Modelling*, 9(3-5): 161–176.

Asano, H.; Mizumoto, T.; and Inui, K. 2017. Reference-Based Metrics Can Be Replaced with Reference-less Metrics in Evaluating Grammatical Error Correction Systems. In Kondrak, G.; and Watanabe, T., eds., *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 343–348. Taipei, Taiwan: Asian Federation of Natural Language Processing.

Awasthi, A.; Sarawagi, S.; Goyal, R.; Ghosh, S.; and Piratla, V. 2019. Parallel Iterative Edit Models for Local Sequence Transduction. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4260–4270. Hong Kong, China: Association for Computational Linguistics.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models Are Few-Shot Learners. https://arxiv.org/abs/2005.14165v4.

Bryant, C.; Felice, M.; and Briscoe, T. 2017. Automatic Annotation and Evaluation of Error Types for Grammatical Error Correction. In Barzilay, R.; and Kan, M.-Y., eds., *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 793–805. Vancouver, Canada: Association for Computational Linguistics.

Choshen, L.; and Abend, O. 2018. Inherent Biases in Reference-based Evaluation for Grammatical Error Correction. In Gurevych, I.; and Miyao, Y., eds., *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 632–642. Melbourne, Australia: Association for Computational Linguistics.

Chu, K.; Chen, Y.-P.; and Nakayama, H. 2024. A Better LLM Evaluator for Text Generation: The Impact of Prompt Output Sequencing and Optimization. arXiv:2406.09972.

Coyne, S.; Sakaguchi, K.; Galvan-Sosa, D.; Zock, M.; and Inui, K. 2023. Analyzing the Performance of GPT-3.5 and GPT-4 in Grammatical Error Correction. arXiv:2303.14342.

Cronbach, L. J. 1951. Coefficient Alpha and the Internal Structure of Tests. *Psychometrika*, 16(3): 297–334.

Dahlmeier, D.; and Ng, H. T. 2012. Better Evaluation for Grammatical Error Correction. In Fosler-Lussier, E.; Riloff, E.; and Bangalore, S., eds., *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 568–572. Montréal, Canada: Association for Computational Linguistics.

Fang, T.; Liu, X.; Wong, D. F.; Zhan, R.; Ding, L.; Chao, L. S.; Tao, D.; and Zhang, M. 2023a. TransGEC: Improving Grammatical Error Correction with Translationese. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023*, 3614–3633. Toronto, Canada: Association for Computational Linguistics.

Fang, T.; Yang, S.; Lan, K.; Wong, D. F.; Hu, J.; Chao, L. S.; and Zhang, Y. 2023b. Is ChatGPT a Highly Fluent Grammatical Error Correction System? A Comprehensive Evaluation. arXiv:2304.01746.

Grundkiewicz, R.; Junczys-Dowmunt, M.; and Gillian, E. 2015. Human Evaluation of Grammatical Error Correction Systems. In Màrquez, L.; Callison-Burch, C.; and Su, J., eds., *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 461–470. Lisbon, Portugal: Association for Computational Linguistics.

Kaneko, M.; Mita, M.; Kiyono, S.; Suzuki, J.; and Inui, K. 2020. Encoder-Decoder Models Can Benefit from Pretrained Masked Language Models in Grammatical Error Correction. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4248–4254. Online: Association for Computational Linguistics.

Kiyono, S.; Suzuki, J.; Mita, M.; Mizumoto, T.; and Inui, K. 2019. An Empirical Study of Incorporating Pseudo Data into Grammatical Error Correction. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 1236–1242. Hong Kong, China: Association for Computational Linguistics.

Kobayashi, M.; Mita, M.; and Komachi, M. 2024. Revisiting Meta-evaluation for Grammatical Error Correction. arXiv:2403.02674.

Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; and Zhu, C. 2023. G-Eval: NLG Evaluation Using GPT-4 with Better Human Alignment. arXiv:2303.16634.

Napoles, C.; Sakaguchi, K.; Post, M.; and Tetreault, J. 2015. Ground Truth for Grammatical Error Correction Metrics. In Zong, C.; and Strube, M., eds., *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)*, 588–593. Beijing, China: Association for Computational Linguistics.

Omelianchuk, K.; Atrasevych, V.; Chernodub, A.; and Skurzhanskyi, O. 2020. GECToR – Grammatical Error Correction: Tag, Not Rewrite. In Burstein, J.; Kochmar, E.; Leacock, C.; Madnani, N.; Pilán, I.; Yannakoudakis, H.; and Zesch, T., eds., *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 163–170. Seattle, WA, USA → Online: Association for Computational Linguistics.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: A Method for Automatic Evaluation of Machine Translation. In Isabelle, P.; Charniak, E.; and Lin, D., eds., *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Rothe, S.; Mallinson, J.; Malmi, E.; Krause, S.; and Severyn, A. 2021. A Simple Recipe for Multilingual Grammatical Error Correction. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 702–707. Online: Association for Computational Linguistics.

Sellam, T.; Das, D.; and Parikh, A. 2020. BLEURT: Learning Robust Metrics for Text Generation. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7881–7892. Online: Association for Computational Linguistics.

Sottana, A.; Liang, B.; Zou, K.; and Yuan, Z. 2023. Evaluation Metrics in the Era of GPT-4: Reliably Evaluating Large Language Models on Sequence to Sequence Tasks. arXiv:2310.13800.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; and Zhou, D. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903.

Wu, H.; Zhang, S.; Zhang, Y.; and Zhao, H. 2023. Rethinking Masked Language Modeling for Chinese Spelling Correction. arXiv:2305.17721.

Yoshimura, R.; Kaneko, M.; Kajiwara, T.; and Komachi, M. 2020. SOME: Reference-less Sub-Metrics Optimized for Manual Evaluations of Grammatical Error Correction. In Scott, D.; Bel, N.; and Zong, C., eds., *Proceedings of the 28th International Conference on Computational Linguistics*, 6516–6522. Barcelona, Spain (Online): International Committee on Computational Linguistics.

Reproducibility Checklist

This paper:

Includes a conceptual outline and/or pseudocode description of AI methods introduced (yes/partial/no/NA)

yes

Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes/no)

yes

Provides well marked pedagogical references for less-familiare readers to gain background necessary to replicate the paper (yes/no)

yes

Does this paper make theoretical contributions? (yes/no)

yes

If yes, please complete the list below.

All assumptions and restrictions are stated clearly and formally. (yes/partial/no)

yes

All novel claims are stated formally (e.g., in theorem statements). (yes/partial/no)

yes

Proofs of all novel claims are included. (yes/partial/no)

yes

Proof sketches or intuitions are given for complex and/or novel results. (yes/partial/no)

yes

Appropriate citations to theoretical tools used are given. (yes/partial/no)

yes

All theoretical claims are demonstrated empirically to hold. (yes/partial/no/NA)

yes

All experimental code used to eliminate or disprove claims is included. (yes/no/NA)

NA

Does this paper rely on one or more datasets? (yes/no)

yes

If yes, please complete the list below. A motivation is given for why the experiments are conducted on the selected datasets (yes/partial/no/NA)

yes

All novel datasets introduced in this paper are included in a data appendix. (yes/partial/no/NA)

NA

All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. (yes/partial/no/NA)

yes

All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations. (yes/no/NA)

yes

All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available. (yes/partial/no/NA)

yes

All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing. (yes/partial/no/NA)

yes

Does this paper include computational experiments? (yes/no)

yes

If yes, please complete the list below.

Any code required for pre-processing data is included in the appendix. (yes/partial/no). All source code required for conducting and analyzing the experiments is included in a code appendix. (yes/partial/no)

no

All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. (yes/partial/no)

yes

All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes/partial/no)

yes

If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results. (yes/partial/no/NA)

yes

This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks. (yes/partial/no)

yes

This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics. (yes/partial/no)

yes

This paper states the number of algorithm runs used to compute each reported result. (yes/no)

yes

Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information. (yes/no)

yes

The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank). (yes/partial/no)

yes

This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments. (yes/partial/no/NA)

yes

This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting. (yes/partial/no/NA)

yes