Don't Count the Edits, Judge by the Outcome Alone: Reward-Based Evaluation for Grammatical Error Correction

Hayeong Ryu[1], Sunhee Jo[2], Seunguk Yu[1], YoungBin Kim[1,2]
[1]Department of Artificial Intelligence, Chung-Ang University
[2]Graduate School of Advanced Imaging Sciences, Multimedia and Film, Chung-Ang University
{bluebarry37, jo3438, bokju128, ybkim85}@cau.ac.kr

Abstract

Grammatical error correction (GEC) evaluation has traditionally relied on reference or edit overlap, which can penalize valid rewrites that differ from gold corrections. Reference-free metrics reduce this dependence, but evaluating whether a fluent output is a valid correction of the source remains challenging. We propose SURE, a source-conditioned reward evaluator trained on within-source preferences spanning minimal-edit and rewrite-oriented corrections. SURE jointly learns an overall reward with criteria-level supervision for grammaticality, faithfulness, and fluency, together with span-level grounding for source-side error resolution. Experiments on SEEDA show that SURE performs competitively against strong baselines, with particular gains on rewrite-style corrections and more disentangled criteria-level diagnostics. Our code is available at https://github.com/hayeonggg/SURE.

1 Introduction

Grammatical error correction (GEC) evaluation has traditionally been defined through reference-based comparison. Metrics such as M^2 (Dahlmeier and Ng, 2012), ERRANT (Ng et al., 2014; Bryant et al., 2017), and GLEU (Napoles et al., 2015, 2019) operationalize this view by comparing a system output against gold corrections using edit- or surface-level overlap. This edit- and reference-centered evaluation paradigm, however, introduces several limitations: it can penalize valid corrections that diverge from the gold reference, under-represent the diversity of acceptable corrections, and favor minimal edits over more fluent rewrites (Napoles et al., 2017; Chollampatt and Ng, 2018; Kobayashi et al., 2024a). Moreover, edit-wise comparison can miss sentence-level properties such as meaning preservation and naturalness, which are central to whether an output is a valid correction of the source (Choshen and Abend, 2018).

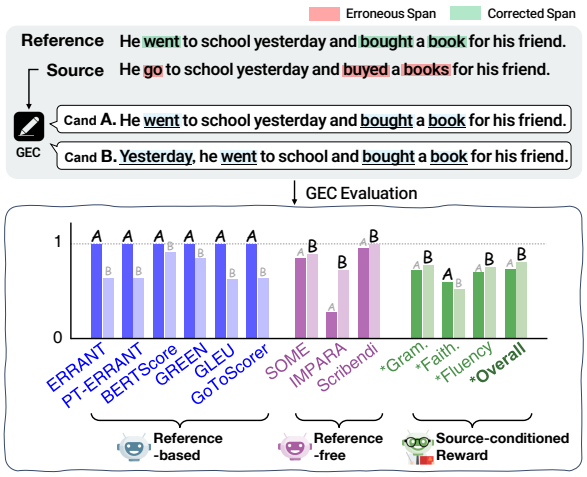


Figure 1: Example of evaluation mismatch between reference-based and reference-free metrics for LLM-style correction. The figure compares two candidate corrections for a source sentence and a reference correction, showing how different metric families assign preferences to minimal-edit and rewrite-style outputs. For SURE, Gram., Faith., Flu., and Overall denote grammaticality, faithfulness, fluency, and overall reward, respectively.

A natural response is to evaluate corrections without relying on a fixed reference. Reference-free and acceptability-based metrics (Asano et al., 2017; Yoshimura et al., 2020; Islam and Magnani, 2021; Maeda et al., 2022; Kobayashi et al., 2024a) shift the focus from reference overlap to the quality of the source–output relation: whether the output corrects the source, preserves its intended meaning, and reads as grammatical and fluent English.

Existing reference-free metrics already condition on both the source and the correction. The remaining issue is therefore not source conditioning itself, but how correction validity is supervised. In particular, prior metrics do not explicitly supervise source-side error resolution or relative preferences among style-diverse corrections of the same source. Thus, the central challenge is not simply to remove

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 6581–6597
October 24-29, 2026 ©2026 Association for Computational Linguistics

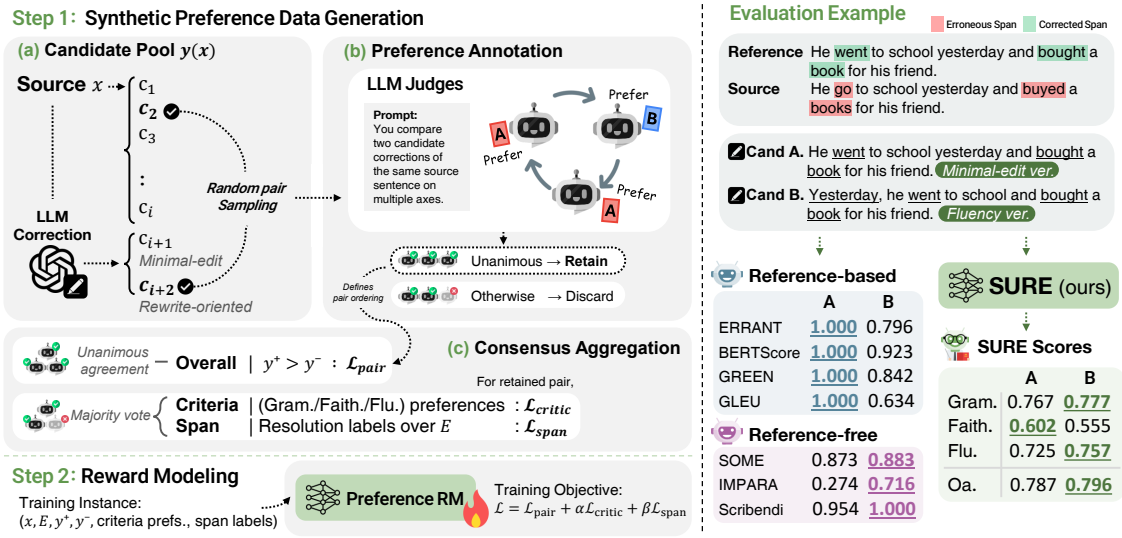


Figure 2: Overview of **SURE**. **(Left)** **Step 1:** Synthetic Preference Data Generation. Given a source sentence x, we construct a style-diverse candidate pool $\mathcal{Y}(x)$ and aggregate LLM-judge annotations into an overall preference, criteria-level preferences, and span-resolution labels. **Step 2:** Reward Modeling. We train a source-conditioned reward model with pairwise, criteria-level, and span-level supervision. **(Right)** We compare reference-based and reference-free metrics with SURE using an illustrative evaluation example. At inference time, SURE scores candidate corrections using only the source and candidate, without references.

references, but to judge whether a fluent output is justified as a correction of the given source.

This issue is amplified in LLM-based correction, where systems often produce holistic rewrites rather than isolated local edits (Fang et al., 2023; Katinskaia and Yangarber, 2024). Such outputs may improve fluency, naturalness, and style while preserving the source meaning, but they can be penalized by reference-based metrics as reference mismatches or over-corrections (Napoles et al., 2017; Kobayashi et al., 2024b). Figure 1 illustrates this mismatch. Candidate A closely follows the gold reference and is therefore strongly favored by reference-based metrics, whereas Candidate B rewrites the sentence into a more fluent form and is preferred by several reference-free metrics. Neither behavior is sufficient on its own: evaluation should not reward reference overlap at the expense of valid rewrites, but it should also not treat fluency alone as evidence of correction quality. This motivates training signals that directly capture whether source-side errors are resolved while accounting for grammaticality, faithfulness, and fluency (Yoshimura et al., 2020; Maeda et al., 2022; Kobayashi et al., 2024a).

This discussion leads to a central question: how can GEC evaluation recognize diverse valid corrections while ensuring that the output remains a faithful correction of the source? We address this question by directly instantiating two requirements of correction validity as training signals: resolving source-side errors and recognizing diverse valid corrections of the same source. Concretely, we combine span-level error-resolution supervision with preferences between minimal-edit and rewrite-oriented corrections.

We propose SURE, a **S**ource-conditioned **U**nified **R**eward-based **E**valuator for GEC. SURE is trained on within-source preference data spanning minimal-edit and rewrite-oriented corrections, and jointly learns an overall reward with criteria-level supervision and span-level grounding for source-side error resolution. Experiments on SEEDA (Kobayashi et al., 2024a) show that SURE aligns strongly with human judgments, especially under +Fluency and rewrite-style comparisons. Criteria-level analysis further shows that SURE provides more disentangled diagnostic signals than prior multi-criteria evaluators.

Our main contributions are as follows:

(1) We introduce a GEC-specific structured supervision design that combines source-side error-resolution signals with within-source preferences over minimal-edit and rewrite-oriented corrections.

(2) We propose SURE, a reference-free reward evaluator that jointly learns an overall reward with criteria-level and span-level supervision

from this preference data.

(3) We show that SURE aligns strongly with human judgments on rewrite-style corrections, and that rewrite-oriented data improves rewrite-related comparisons while introducing a trade-off on traditional GEC pairs.

2 Problem Formulation

Task Formulation. Conventional GEC metrics typically score y by comparing it with a reference correction z, often through edit or surface-form overlap (Dahlmeier and Ng, 2012; Napoles et al., 2015; Bryant et al., 2017; Napoles et al., 2019). For SURE, we adopt a source-conditioned reward formulation, where an evaluator assigns a scalar reward $R_\theta(x, y)$ that reflects the validity of y as a correction of x. This formulation enables reference-free evaluation at inference time, with correction validity as the scoring target.

Correction Validity Criteria. Although GEC has traditionally emphasized minimal corrections, the minimal edit is not always the best correction when meaning-preserving rewrites can improve naturalness (Napoles et al., 2017). Prior work has therefore evaluated GEC quality as a multi-dimensional judgment beyond grammatical correctness alone (Asano et al., 2017; Choshen and Abend, 2018; Yoshimura et al., 2020; Maeda et al., 2022). Following this perspective, we define correction validity using three criteria: (1) **Grammaticality** measures whether the candidate is grammatically well-formed and resolves source-side errors without introducing new ones; (2) **Faithfulness** measures whether the candidate preserves the meaning and intent of the source sentence; and (3) **Fluency** measures whether the corrected sentence is natural and idiomatic beyond being merely grammatical. These criteria provide the basis for how we characterize correction validity in this work.

3 SURE: Source-conditioned Unified Reward-based Evaluator

We introduce SURE, a reference-free reward evaluator for GEC that scores whether a candidate is a valid correction of the source sentence. Figure 2 provides an overview of the proposed framework, which consists of two stages: synthetic preference data generation and reward modeling. We first describe the construction of synthetic preference data from style-diverse correction candidates and consensus among LLM judges. Then we define a reward model that predicts an overall correction reward and criteria-level scores.

3.1 Synthetic Preference Data Generation

Candidate Pool Construction. We first collect source sentences from three standard GEC benchmarks: BEA-2019 (Bryant et al., 2019), CoNLL-2014 (Ng et al., 2014), and JFLEG (Napoles et al., 2017). We exclude all CoNLL-2014-derived data to avoid source overlap with SEEDA. To focus on examples that require non-trivial correction judgments, we restrict the source set to sentences with at least 10 tokens and at least two errors identified by ERRANT (Ng et al., 2014; Bryant et al., 2017). We further use ERRANT to identify source-side error spans from source–reference pairs, which serve as anchors for span-level annotation.[1]. Each candidate pool consists of existing candidates, including human references and GEC system outputs, and two additional corrections generated using GPT-4o (OpenAI, 2024). The GPT-4o-based candidates are generated with two different instructions: one for *minimal-edit correction* and the other for *rewrite-oriented correction*, which encourages a fluent rewrite with a different surface form while preserving the original meaning. Including rewrite-oriented corrections exposes the preference data to valid corrections that differ substantially from minimal edits.[2] The prompts used for LLM-generated corrections are provided in Appendix D.

Preference Annotation. To construct pairwise preference data, we evaluate correction candidate pairs for each source sentence. Candidate pairs are randomly sampled from each source's candidate pool and evaluated by three independent LLM judges: GPT-4.1-mini (OpenAI, 2025), Claude Haiku 4.5 (Anthropic, 2025), and Grok 4.3 (xAI, 2026). For each comparison, the judges select the better correction overall, assess the candidates along the criteria defined in Section 2, namely grammaticality, faithfulness, and fluency, and annotate whether each marked source error is resolved by each candidate. We keep only pairs with unanimous agreement on the overall preference, while aggregating criteria-level preferences and span labels by majority vote. This process

[1] References serve two roles during data construction: they are included as candidate corrections and are used in source–reference pairs to localize source-side errors for span supervision. At inference time, SURE does not require references.

[2] The impact of rewrite-oriented corrections is analyzed in Section 5.2

Metric	System-level								Sentence-level							
	SEEDA-E				SEEDA-S				SEEDA-E				SEEDA-S			
	Base		+Fluency		Base		+Fluency		Base		+Fluency		Base		+Fluency	
	r	ρ	r	ρ	r	ρ	r	ρ	Acc.	τ	Acc.	τ	Acc.	τ	Acc.	τ
Reference-based																
ERRANT	0.386	0.364	-0.576	-0.130	0.242	0.084	-0.656	-0.305	0.629	0.257	0.568	0.135	0.588	0.177	0.517	0.034
PT-ERRANT	0.588	0.545	-0.536	0.024	0.464	0.273	-0.617	-0.147	0.613	0.226	0.554	0.109	0.591	0.181	0.517	0.034
BERTScore	-0.031	0.552	-0.513	-0.002	-0.046	0.294	-0.549	-0.165	0.563	0.125	0.500	-0.001	0.558	0.115	0.477	-0.046
GREEN	0.002	0.657	0.017	0.349	0.016	0.580	0.013	0.301	0.600	0.201	0.557	0.114	0.599	0.198	0.546	0.093
GLEU	0.224	0.650	0.526	0.767	0.243	0.608	0.542	0.741	0.655	0.309	0.638	0.276	0.662	0.325	0.630	0.260
GoToScorer	0.843	0.881	-0.410	0.323	0.722	0.706	-0.503	0.191	0.704	0.408	0.644	0.288	0.675	0.350	0.603	0.207
CLEME 2.0	0.643	0.650	0.628	0.705	0.656	0.587	0.598	0.666	0.713	0.427	0.670	0.339	0.730	0.461	0.657	0.314
JELV 2.0	0.967	0.944	0.946	0.965	0.883	0.811	0.938	0.881	0.776	0.552	0.757	0.514	0.741	0.482	0.722	0.444
Reference-free																
SOME	0.901	0.951	0.943	0.969	0.892	0.867	0.931	0.916	0.771	0.542	0.758	0.516	0.783	0.565	0.769	0.537
Scribendi	0.825	0.839	0.715	0.842	0.620	0.636	0.604	0.714	0.760	0.519	0.742	0.484	0.740	0.480	0.707	0.414
IMPARA	0.902	0.965	0.900	0.978	0.916	**0.902**	0.887	**0.938**	0.757	0.514	0.747	0.493	0.758	0.516	0.741	0.481
GPT-4.1-E	0.584	0.602	0.893	0.746	0.520	0.634	0.874	0.766	0.741	0.482	0.742	0.483	0.733	0.465	0.747	0.494
GPT-4.1-S	0.203	0.266	0.772	0.534	0.080	0.273	0.745	0.538	0.739	0.478	0.741	0.482	0.736	0.472	0.748	0.496
GPT-4-S†	0.960	0.958	0.967	0.969	0.887	0.860	0.931	0.908	0.798	0.595	0.783	0.565	0.784	0.567	0.770	0.540
+ Fluency †	**0.974**	**0.979**	**0.981**	**0.982**	0.913	0.874	0.952	0.916	**0.831**	**0.662**	**0.812**	**0.624**	**0.819**	**0.637**	0.797	0.594
SURE (ours)	0.932	0.972	0.976	**0.982**	**0.927**	0.895	**0.970**	0.934	0.797	0.595	0.796	0.591	0.809	0.619	**0.808**	**0.615**

Table 1: Meta-evaluation results on SEEDA. Results are reported for both system-level and sentence-level evaluation under the Base and +Fluency settings. Metrics are grouped by whether they require references at inference time. Higher values indicate better agreement with human judgments; the best score is in bold and the second-best score is underlined. † indicates scores reported in (Kobayashi et al., 2024a), rather than results rerun in our experimental setup.

yields 2,400 preference pairs from 1,320 unique source sentences. Filtering statistics are provided in Appendix A, and we validate the filtering strategy with human judgments in Section 5.4.

3.2 Reward Modeling

Model Architecture. We train SURE as a source-conditioned reward model using the preference data constructed in Section 3.1. Given a source sentence x and a candidate correction y, the model encodes the concatenated input $[x; y]$ and predicts an overall reward together with three auxiliary criterion scores $R_\theta^c(x, y)$ for each $c \in \{\text{gram}, \text{faith}, \text{flu}\}$. We use DeBERTa-v3-large (He et al., 2023) as the encoder and fine-tune it with LoRA (Hu et al., 2022). The overall head is used as the main reward signal, while the criterion heads provide auxiliary supervision for grammaticality, faithfulness, and fluency.

Training Objective. For each preference pair (x, y^+, y^-), we optimize the model to score the preferred correction y^+ higher than the dispreferred correction y^-. The training objective combines pairwise reward learning, criteria-level supervision, and span-level grounding:

$$\mathcal{L} = \mathcal{L}_{\text{pair}} + \alpha \mathcal{L}_{\text{critic}} + \beta \mathcal{L}_{\text{span}}, \quad (1)$$

where $\mathcal{L}_{\text{pair}}$ denotes the pairwise ranking loss over the overall reward, and $\mathcal{L}_{\text{critic}}$ denotes criteria-level pairwise ranking losses for grammaticality, faithfulness, and fluency. Each criterion-specific loss uses its corresponding preference label, while $\mathcal{L}_{\text{span}}$ is an auxiliary binary classification loss for predicting whether each marked error span in the source is resolved by the candidate correction[3]. The weights α and β are hyperparameters that balance criteria-level and span-level supervision, respectively.

Span supervision is used only during training: when computing $\mathcal{L}_{\text{span}}$, we mark error spans in the source so that the encoder learns whether the candidate resolves the corresponding local errors. At inference time, the span markers and span head are removed, and SURE takes only the raw source sentence and candidate correction as input. This allows the model to remain reference-free at evaluation time while using span-level signals to reduce its reliance on surface fluency alone.

[3]Span annotations are binarized by mapping *resolved* to positive and both *missed* and *partial* to negative, followed by majority voting across judges.

Metric	SEEDA-E				SEEDA-S			
	All	**R–T**	**R–R**	**T–T**	**All**	**R–T**	**R–R**	**T–T**
ERRANT	48.2	39.8	32.4	53.8	49.8	42.0	35.8	55.7
BERTScore	41.0	30.4	29.6	47.4	43.8	35.9	35.5	49.4
GLEU	51.9	52.7	40.9	51.6	55.4	57.3	41.9	54.4
SOME	60.1	67.1	53.5	55.6	62.4	68.3	45.0	58.5
IMPARA	58.9	63.8	53.5	55.7	59.9	63.8	53.3	57.1
GPT-4.1-E	60.2	71.9	53.3	54.7	59.8	69.9	52.3	54.4
SURE	**65.4**	**73.3**	**58.1**	**60.4**	**65.7**	**72.3**	**58.3**	**61.0**

Table 2: Pairwise accuracy for rewrite-style correction evaluation. Comparisons are grouped by system style: **R–T** compares rewrite-style and traditional GEC systems, **R–R** compares two rewrite-style systems, and **T–T** compares two traditional GEC systems; **All** includes all pairs. Values are the percentage of metric preferences matching human preferences, with the best score in each column in **bold**.

Metric	Setting	SEEDA-E		SEEDA-S	
		Base	+Fluency	Base	+Fluency
System –level (r)	w/o rewrite	0.910	0.967	0.922	0.962
	w/ rewrite	0.926	0.978	0.945	0.989
		(+0.016)	(+0.011)	(+0.023)	(+0.027)
Sentence –level (Acc.)	w/o rewrite	0.808	0.802	0.820	0.821
	w/ rewrite	0.791	0.785	0.822	0.821
		(-0.017)	(-0.017)	(+0.002)	(+0.000)

Table 3: Effect of rewrite-oriented preference data. w/o rewrite denotes training without rewrite candidates, while w/ rewrite denotes training with rewrite candidates. Values in parentheses indicate the difference from w/o rewrite. Ablation models are retrained under a controlled setting, so their absolute scores are not intended to exactly match the final SURE scores in Table 1; we focus on relative differences between ablation conditions.

4 Experiments

4.1 Experimental Setup

Meta-Evaluation. We conduct our main evaluation on SEEDA (Kobayashi et al., 2024a), a meta-evaluation benchmark for GEC metrics. SEEDA consists of two subsets: SEEDA-E for edit-level pairwise human judgments and SEEDA-S for sentence-level pairwise human judgments. Both subsets cover corrections from twelve GEC systems in the Base setting, with two additional fluent human corrections introduced in the +Fluency setting. In this setting, the added fluent corrections test whether a metric can recognize rewrite-style corrections as valid when they preserve meaning and improve naturalness, even if they differ substantially from reference edits.

At the sentence level, we convert metric scores into pairwise preferences and compare them with human judgments using accuracy and Kendall's τ. For system-level evaluation, we aggregate metric scores over corrections for each system and compare the resulting ranking with the human ranking using Pearson's r and Spearman's ρ.

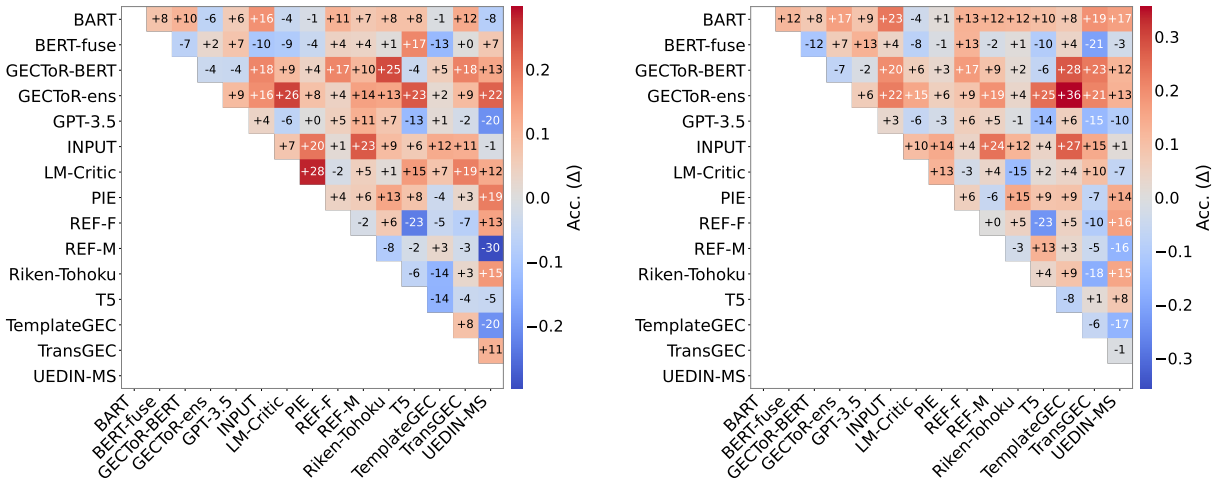
Baselines. We compare SURE against two groups of GEC evaluation baselines: reference-based metrics and reference-free metrics. The **reference-based metrics** include ERRANT (Ng et al., 2014; Bryant et al., 2017), PT-ERRANT (Gong et al., 2022), BERTScore (Zhang et al., 2020), GREEN (Koyama et al., 2024), GLEU (Napoles et al., 2015, 2019), GoTo-Scorer (Gotou et al., 2020), CLEME 2.0 (Ye et al., 2025), and JELV 2.0 (Zhan et al., 2026). The **reference-free metrics** include SOME (Yoshimura et al., 2020), Scribendi (Islam and Magnani, 2021), IMPARA (Maeda et al., 2022), and LLM-{S,E} evaluators (Kobayashi et al., 2024a). For LLM-{S,E}, we instantiate the evaluator with GPT-4.1-mini[4] and report the edit-level and sentence-level variants as GPT-4.1-E and GPT-4.1-S, respectively. We additionally include the reported GPT-4-S (+Fluency) results (Kobayashi et al., 2024a), obtained using GPT-4 (OpenAI, 2023) (`gpt-4-1106-preview`). All non-LLM baselines are evaluated using `gec-metrics`, a unified evaluation framework for GEC metrics (Goto et al., 2025).

Implementation. We implement the model with LoRA using rank 16, scaling factor 32, and dropout 0.1. The maximum sequence length is set to 256. We train for 5 epochs with learning rate 2×10^{-4}, weight decay 0.01, warmup ratio 0.1, gradient clipping 1.0, and an effective batch size of 8 via gradient accumulation. The loss weights are set to $\alpha = 0.5$ and $\beta = 0.2$.

4.2 Correction Validity Evaluation

We evaluate each metric against human judgments on SEEDA. As shown in Table 1, reference-based metrics perform reasonably in the Base setting, but their agreement often drops under +Fluency, where valid corrections may diverge from reference edits. Reference-free metrics are generally more stable in this setting, indicating the advantage of evaluating

[4]https://openai.com/index/gpt-4-1/



(a) SEEDA-E (b) SEEDA-S

Figure 3: System-pair accuracy difference between SURE and GPT-4.1-E. We select GPT-4.1-E as the comparison system because it is the reference-free LLM evaluator most directly comparable to SURE, assessing edit appropriateness conditioned on both the source and the candidate. Each cell reports the pairwise accuracy difference, computed as $\text{Acc.}_{\text{SURE}} - \text{Acc.}_{\text{GPT-4.1-E}}$. Positive values indicate higher agreement with human preferences by SURE, while negative values indicate higher agreement by GPT-4.1-E.

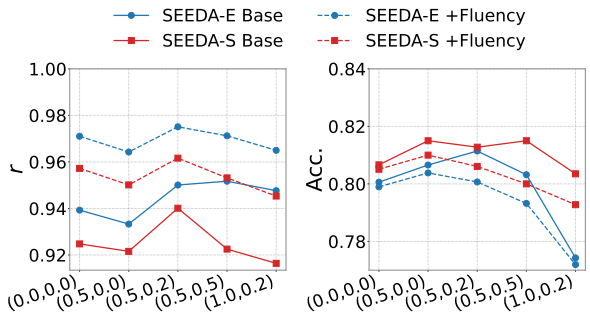


Figure 4: Sensitivity to auxiliary loss weights α and β in Eq. 1. We report system-level Pearson's r and sentence-level accuracy (Acc.) on SEEDA-E and SEEDA-S under Base and +Fluency settings. The default setting is $(\alpha, \beta) = (0.5, 0.2)$.

Dataset	Pair Type	w/o rewrite	w/ rewrite
SEEDA-E	All	67.30	66.68 (-0.62pp)
	Trad. only	61.03	58.77 (-2.26pp)
	Rewrite-inv.	76.50	78.30 (+1.80pp)
	Rewrite vs. Trad.	77.15	79.20 (+2.05pp)
SEEDA-S	All	66.52	65.45 (-1.07pp)
	Trad. only	61.07	57.39 (-3.68pp)
	Rewrite-inv.	73.41	75.64 (+2.23pp)
	Rewrite vs. Trad.	74.26	76.37 (+2.11pp)

Table 4: Pair-type analysis of rewrite-oriented preference data. w/o rewrite denotes training without rewrite candidates, while w/ rewrite denotes training with rewrite candidates. Trad. only compares two traditional GEC systems, Rewrite-inv. includes pairs where at least one system is rewrite-style, and Rewrite vs. Trad. compares a rewrite-style system with a traditional GEC system. Values are pairwise accuracy (%) against human preferences.

correction validity without relying on reference overlap. Among reference-free metrics, SURE maintains strong agreement across both Base and +Fluency settings at the system and sentence levels. The results indicate that source-conditioned reward estimation provides a robust criterion for evaluating corrections beyond reference overlap. Additional results on cross-domain transfer are provided in Appendix C.1.

5 Analysis

5.1 Evaluation Behavior of SURE

Rewrite-Style Correction Evaluation. We further group SEEDA pairs by comparison type and evaluate which metric best matches human preferences within each group, as shown in Table 2. SURE achieves the highest accuracy across all groups on both SEEDA-E and SEEDA-S. The separation is most visible in rewrite-related comparisons, where edit-overlap metrics fall behind reference-free metrics. This result shows that SURE is better aligned with human judgments when evaluating rewrite-style corrections, while remaining competitive on traditional GEC comparisons.

System-Pair Analysis. We further compare SURE with GPT-4.1-E at the system-pair level,

Judge/ Aggregator	Overall		Gram.		Faith.		Flu.	
	Acc.	κ	Acc.	κ	Acc.	κ	Acc.	κ
GPT-4.1-mini	0.760	0.515	0.770	0.536	0.395	-0.200	0.770	0.533
Claude Haiku 4.5	0.805	0.606	0.765	0.527	**0.685**	**0.366**	0.820	0.635
Grok 4.3	0.755	0.509	0.720	0.440	0.470	-0.052	0.805	0.607
Majority	0.790	0.577	0.745	0.487	0.490	-0.018	0.810	0.614
Unanimous	**0.831**	**0.658**	**0.819**	**0.635**	0.561	0.129	**0.842**	**0.677**

Table 5: Reliability of LLM-based preference annotations on SEEDA pairs. The first three rows report individual judge performance. Majority aggregates the three LLM judges by majority vote, while unanimous retains only pairs on which all judges agree.

as shown in Figure 3. Positive differences appear across many system pairs, indicating that SURE's advantage is not limited to aggregate scores. The gains are especially frequent around rewrite-style systems such as GPT-3.5 and REF-F, while negative cells remain localized to a small number of specific pairs.

Auxiliary supervision. We analyze whether the auxiliary criteria-level and span-level losses contribute to the overall reward model, varying their loss weights as summarized in Figure 4. The setting $(0.0, 0.0)$ corresponds to the pairwise-only model, while $(0.5, 0.0)$ removes span-level supervision and isolates the effect of criteria-level learning. The results show that SURE is not highly sensitive to the exact choice of these weights, with stable performance across both system- and sentence-level evaluations. At the same time, using auxiliary supervision generally improves or preserves performance compared to the pairwise-only objective, indicating that criteria- and span-level signals provide useful grounding for the overall reward.

We further examine the system-level consequences of these evaluation mismatches under the +Fluency setting; complete ranking analyses are provided in Appendix B.

5.2 Effect and Reliability of Preference Data

Effect of Rewrite-Oriented Data. We ablate rewrite-oriented candidates during preference-data construction to examine their contribution to SURE (Table 3). Adding rewrite data consistently improves system-level correlation across both SEEDA-E and SEEDA-S, with larger gains in the +Fluency setting. This indicates that style-diverse preference data helps the model rank systems more accurately when fluent rewrite-style corrections are included. Sentence-level accuracy is largely preserved, although SEEDA-E shows a small decrease, suggesting a mild trade-off with edit-level judgments that favor more local corrections.

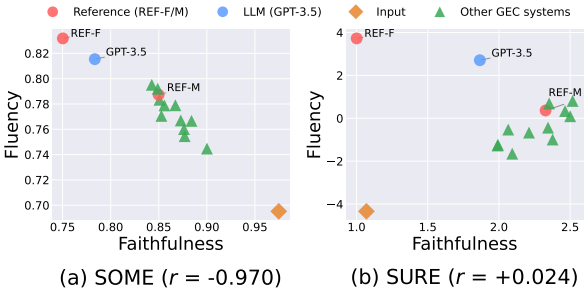


Figure 5: Criteria-level diagnostics of faithfulness and fluency. Each point represents a GEC system, with scores averaged over its outputs. For SOME, the meaning-preservation score is treated as faithfulness for consistency. Pearson's r measures the correlation between the two criteria.

Pair-Type Effects of Rewrite-Oriented Data. Table 4 further breaks down sentence-level accuracy by system style. Comparing models trained with and without rewrite candidates, we find that rewrite-oriented data improves pairs involving rewrite-style systems, especially Rewrite vs. Trad. comparisons, but reduces accuracy on Trad. only pairs. This explains the small change in aggregate sentence-level accuracy: gains on rewrite-related comparisons are partially offset by drops on traditional-only comparisons. Thus, rewrite-oriented data mainly improves the evaluation of rewrite-style systems, while introducing a mild trade-off for minimal-edit-oriented comparisons.

Reliability of Consensus-Based Preference Data. The quality of preference data depends on whether LLM judges provide judgments that are consistent with human preferences. Table 5 shows that enforcing unanimous agreement improves reliability over both individual judges and majority voting, yielding the highest overall accuracy and κ. This provides empirical support for using unanimous consensus as the filtering criterion for constructing training pairs.

5.3 Criteria-Level Diagnostics

We evaluate whether the criteria-level scores offer diagnostic information beyond the overall reward. A useful multi-criteria evaluator should distinguish different correction profiles, such as fluency-oriented rewrites and meaning-preserving conservative corrections. As shown in Figure 5,

Metric	Retained (3–0)	Discarded (2–1)
A/B consensus	62.0% (31/50)	38.0% (19/50)
Human pairwise agreement	49.7%	41.7%
LLM–Human (Overall)	77.4% (24/31)	57.9% (11/19)

Table 6: Human validation of retained and discarded preference pairs. A/B consensus denotes the proportion of pairs for which human annotators reached consensus, and Human pairwise agreement is the raw percent agreement across all six annotator pairs on the three-way Overall judgment (A, B, or Difficult). LLM–Human (Overall) measures agreement between the LLM preference and the human consensus judgment. Retained pairs received unanimous 3–0 LLM judgments, whereas discarded pairs received 2–1 split judgments.

Criterion	Agreement	Cases
Overall	77.4%	24/31
Gram.	92.3%	12/13
Faith.	84.0%	21/25
Flu.	75.9%	22/29

Table 7: LLM–Human agreement on retained preference pairs. Agreement is computed over pairs with human consensus for each dimension. Cases report the number of LLM–Human agreements over the number of human-consensus pairs.

SOME (Yoshimura et al., 2020) exhibits a strong negative correlation between faithfulness and fluency, indicating that the two scores largely behave as opposite ends of a single axis. By contrast, SURE yields nearly independent faithfulness and fluency scores, allowing these correction profiles to be separated in the criteria space. A similar separation is observed between grammaticality and faithfulness, while grammaticality and fluency remain closely related (see Appendix C.3). In Appendix C.2, we further evaluate criteria-level transfer on TMU-GFM; SURE shows reasonable zero-shot transfer on confident grammaticality and fluency pairs, while meaning preservation remains challenging.

5.4 Human Validation of Preference Data

Human Evaluation Setup. To assess whether the LLM-generated preference labels align with independent human judgments, we conduct a human evaluation with four annotators on 100 candidate pairs: 50 retained pairs with 3–0 unanimous LLM agreement and 50 discarded pairs with 2–1 split judgments. The pairs are approximately evenly distributed across the three pair types. Human consensus is defined as a strict majority of at least three of four valid votes.

Measure	Value
Human rewrite preference	83.3%
LLM rewrite preference	66.7%
Human agreement with LLM rewrite	87.5% (7/8)

Table 8: Rewrite preference analysis on human-consensus rewrite–non-rewrite pairs. The first two rows report the rewrite preference rates of human annotators and LLM judges, respectively. The last row reports human agreement among cases in which the LLM judges preferred the rewrite.

Effect of Unanimous Filtering. As shown in Table 6, retained pairs exhibit higher human consensus, pairwise agreement, and LLM–Human agreement than discarded pairs. In particular, the human consensus rate increases from 38.0% for discarded pairs to 62.0% for retained pairs, and the difference is statistically significant ($p = 0.027$). Among pairs for which humans reach consensus, the retained LLM labels agree with the human overall preference in 77.4% of cases, compared with 57.9% for discarded pairs. These results suggest that unanimous filtering preferentially retains less ambiguous comparisons and improves the human alignment of the preference labels used for training.

Criteria-Level Agreement. We further compare the retained LLM labels with human consensus separately for each evaluation criterion. Table 7 shows that agreement is highest for grammaticality (92.3%), followed by faithfulness (84.0%) and fluency (75.9%). Together with the 77.4% overall agreement, these results indicate that the supervision used to train SURE is broadly consistent with independent human judgments across both the overall and criteria-level labels.

Audit of Rewrite Preference Bias. We further assess whether LLM-judge supervision introduces a systematic preference for rewrite-style corrections. As shown in Table 8, among rewrite–non-rewrite pairs with human consensus, the rewrite preference rate is higher for human annotators than for the LLM judges (83.3% vs. 66.7%). Furthermore, human judgments agree with seven of the eight cases in which the LLM judges prefer the rewrite. These results suggest that the observed rewrite preferences are not driven by a systematic over-preference for rewrites by the LLM judges.

6 Related Work

6.1 Reference-Based Metrics for GEC

Reference-based metrics have long served as the dominant framework for GEC evaluation. M^2 and ERRANT compare extracted system edits against gold edits (Dahlmeier and Ng, 2012; Bryant et al., 2017). GLEU evaluates corrections through n-gram overlap with references and the source sentence (Napoles et al., 2015). Subsequent metrics refine this paradigm through contextual or pretrained scoring, as in BERTScore and PT-ERRANT (Zhang et al., 2020; Gong et al., 2022); alignment-free comparison, as in GREEN (Koyama et al., 2024); or difficulty-aware edit weighting, as in GoToScorer (Gotou et al., 2020).

Recent work has further improved interpretability and validity within the edit-based framework. CLEME 2.0 decomposes edit outcomes into hit-, wrong-, under-, and over-correction (Ye et al., 2025), while JELV introduces an edit-level validity judge for evaluation and reference expansion (Zhan et al., 2026). Meta-evaluation studies have also shown that conclusions from reference-based metrics can depend on system sets, human-evaluation granularity, and reference coverage (Chollampatt and Ng, 2018; Kobayashi et al., 2024a). These studies make reference-based evaluation more robust and interpretable, but the target remains tied to gold corrections or reference-derived edits. We evaluate whether a candidate is a valid correction of the source without requiring reference overlap at inference time.

6.2 Reference-Free and LLM-Based GEC Evaluation

Reference-free metrics evaluate GEC outputs without comparing them against gold corrections, reducing dependence on reference coverage. SOME decomposes reference-free evaluation into grammaticality, fluency, and meaning preservation (Yoshimura et al., 2020), while Scribendi proposes a straightforward reference-free alternative to gold-standard comparison (Islam and Magnani, 2021). IMPARA estimates correction quality from the impact of a candidate correction on the source sentence using parallel grammatical and ungrammatical sentence pairs (Maeda et al., 2022).

These approaches move beyond surface overlap and already condition evaluation on the source and candidate correction. However, they do not explicitly combine source-side error-resolution supervision with relative preferences among style-diverse corrections of the same source. Recent work has also explored LLMs as reference-free evaluators for GEC (Kobayashi et al., 2024a), using edit-level and sentence-level prompts and showing that LLMs can capture broader aspects of correction quality than overlap-based metrics. Our work instead focuses on this supervision design, combining source-side error-resolution signals with preferences among minimal-edit and rewrite-oriented corrections.

6.3 Preference and Reward Modeling for Evaluation

Preference-based reward modeling learns evaluators from pairwise judgments by assigning higher scores to preferred outputs, a formulation widely used in language generation and alignment (Christiano et al., 2017; Ziegler et al., 2020; Stiennon et al., 2020; Ouyang et al., 2022; Bai et al., 2022; Rafailov et al., 2023). Learned text generation metrics such as BLEURT, COMET, BARTScore, and UniEval further capture semantic and quality-oriented judgments beyond surface overlap (Sellam et al., 2020; Rei et al., 2020; Yuan et al., 2021; Zhong et al., 2022).

Recent work uses LLMs as judges to collect preferences or directly evaluate generated outputs (Liu et al., 2023; Wang et al., 2023; Zheng et al., 2023; Chen et al., 2024; Fu et al., 2024; Dubois et al., 2024). While such approaches provide scalable supervision, their judgments can vary with prompts, judge models, and evaluation criteria (Zheng et al., 2023; Liu et al., 2023; Lambert et al., 2025). For correction evaluation, preferences should reflect whether an output is justified as a correction of the given source, rather than whether it is merely fluent or well-written. We adopt preference-based evaluation in this source-conditioned setting, assessing correction quality through grammaticality, faithfulness, fluency, and span-level error resolution.

7 Conclusion

We presented SURE, a source-conditioned reward evaluator for grammatical error correction. SURE learns correction validity from style-diverse preferences with criteria-level and span-level supervision. Experiments on SEEDA show strong alignment with human judgments, particularly for fluent rewrite-style corrections, while providing useful diagnostics for grammaticality, faithfulness, and fluency.

Limitations

This work has several limitations. First, SURE is trained on synthetic and LLM-annotated preference data, and may inherit biases from the judge models used during data construction. Second, SURE exhibits uneven cross-domain generalization, with weaker system-level correlations on Wiki despite competitive transfer on FCE and at the sentence level. This suggests that its calibration may remain sensitive to domain shift, motivating broader validation across domains and languages. Third, although references are not required at inference time, reference-derived ERRANT spans are used during training for span-level supervision. Future work will expand human-annotated preference data, evaluate multilingual settings, and improve the calibration of criteria-level rewards.

Acknowledgments

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [RS-2021-II211341, Artificial Intelligence Graduate School Program (Chung-Ang University)] and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-00556246).

References

Anthropic. 2025. Claude Haiku 4.5 system card. https://www.anthropic.com/claude-haiku-4-5-system-card. Accessed: 2026-05-25.

Hiroki Asano, Tomoya Mizumoto, and Kentaro Inui. 2017. Reference-based metrics can be replaced with reference-less metrics in evaluating grammatical error correction systems. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 343–348, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, and 12 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *Preprint*, arXiv:2204.05862.

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.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. 2024. AlpaGasus: Training a better alpaca with fewer data. In *The Twelfth International Conference on Learning Representations*.

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. 2018. Referenceless measure of faithfulness for grammatical error correction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 124–129, New Orleans, Louisiana. Association for Computational Linguistics.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, pages 4302–4310.

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.

Yann Dubois, Percy Liang, and Tatsunori B. Hashimoto. 2024. Length-controlled AlpacaEval: A simple debiasing of automatic evaluators. In *First Conference on Language Modeling*.

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.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. GPTScore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 6556–6576, Mexico City, Mexico. Association for Computational Linguistics.

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 Goto, Yusuke Sakai, and Taro Watanabe. 2025. gec-metrics: A unified library for grammatical error correction evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 524–534, Vienna, Austria. 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.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

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.

Anisia Katinskaia and Roman Yangarber. 2024. GPT-3.5 for grammatical error correction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7831–7843, Torino, Italia. ELRA and ICCL.

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.

Shota Koyama, Ryo Nagata, Hiroya Takamura, and Naoaki Okazaki. 2024. n-gram F-score for evaluating grammatical error correction. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 303–313, Tokyo, Japan. Association for Computational Linguistics.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. 2025. RewardBench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1755–1797, Albuquerque, New Mexico. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using Gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

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, Maria Nădejde, and Joel Tetreault. 2019. Enabling robust grammatical error correction in new domains: Data sets, metrics, and analyses. *Transactions of the Association for Computational Linguistics*, 7:551–566.

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. 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.

OpenAI. 2023. GPT-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI. 2024. GPT-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI. 2025. Introducing GPT-4.1 in the API. https://openai.com/index/gpt-4-1/.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is ChatGPT a good NLG evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Singapore. Association for Computational Linguistics.

xAI. 2026. Grok 4.3. https://docs.x.ai/developers/models/grok-4.3. Accessed: 2026-05-25.

Jingheng Ye, Zishan Xu, Yinghui Li, Linlin Song, Qingyu Zhou, Hai-Tao Zheng, Ying Shen, Wenhao Jiang, Hong-Gee Kim, Ruitong Liu, Xin Su, and Zifei Shan. 2025. CLEME2.0: Towards interpretable evaluation by disentangling edits for grammatical error correction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 204–222, Vienna, Austria. Association for Computational Linguistics.

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.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTScore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277.

Yuhao Zhan, Yuqing Zhang, Jing Yuan, Qixiang Ma, Zhiqi Yang, Yu Gu, Zemin Liu, and Fei Wu. 2026. JELV: A judge of edit-level validity for evaluation and automated reference expansion in grammatical error correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 34611–34619.

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. Fine-tuning language models from human preferences. *Preprint*, arXiv:1909.08593.

A Dataset Statistics

Dataset Composition. The final dataset contains 2,400 preference pairs from 1,320 unique source sentences, drawn from BEA-2019 and JFLEG (Table 9). CoNLL-2014-derived data are excluded to avoid source overlap with SEEDA.

Dataset	Preference pairs	Unique sources
BEA-2019	1,500	881
JFLEG	900	439
Total	2,400	1,320

Table 9: Dataset composition by source benchmark.

Preference Filtering. The three LLM judges evaluated 8,093 candidate pairs, of which 4,846 (59.9%) reached unanimous agreement on the overall preference, while 3,247 received 2–1 split judgments. From these unanimous pairs, we randomly sampled 2,400 pairs to construct the final training set.

Candidate Source Coverage. GPT-4o-generated corrections appear in 87.0% of all preference pairs, accounting for 62.5% of preferred candidates and 40.4% of dispreferred candidates (Table 10). This indicates that the training data provides substantial supervision over LLM-generated corrections, including rewrite-style corrections. The minimal-edit GPT-4o variant has a win rate of 53.8%, while the rewrite-oriented GPT-4o variant has a higher win rate of 69.6%. This difference suggests that rewrite-oriented corrections are frequently preferred when they provide valid and fluent alternatives, which we further examine in Section 5.2.

Statistic	Value
Pairs containing at least one GPT-4o correction	87.0%
Preferred candidates from GPT-4o corrections	62.5%
Dispreferred candidates from GPT-4o corrections	40.4%
Minimal-edit GPT-4o variant win rate	53.8%
Rewrite-oriented GPT-4o variant win rate	69.6%

Table 10: Coverage and preference rates of GPT-4o-generated corrections.

Style Pair Distribution. Table 11 reports the distribution of preference pairs according to the correction styles of the preferred and dispreferred candidates. The arrow denotes the preference direction, with the style on the left corresponding to the preferred correction and the style on the right to the dispreferred correction. The dataset contains substantial comparisons between rewrite-style and minimal-edit corrections, as well as comparisons involving fluency-oriented and rewrite-style corrections.

Style pair ($y^+ \rightarrow y^-$)	Count
rewrite → minimal edit	601
minimal edit → rewrite	560
rewrite → fluency edit	518
rewrite → rewrite	382
fluency edit → fluency edit	311
fluency edit → rewrite	28
Total	2,400

Table 11: Distribution of preference pairs by correction style.

Criteria Labels and Instance Format. For each criterion, Table 12 reports whether the overall preferred correction y^+ or the dispreferred correction y^- is preferred under that criterion. Each training instance contains a source sentence, ERRANT-identified error spans, a preferred correction y^+, a dispreferred correction y^-, criteria-level preferences, span-resolution labels for both candidates, and metadata.

Criterion	y^+ preferred	y^- preferred
Grammaticality	2,367	33
Faithfulness	1,264	1,136
Fluency	2,179	221

Table 12: Alignment between criteria-level preferences and the overall preference.

B System-Ranking Analysis under +Fluency

To examine whether the evaluation mismatch observed for rewrite-style corrections affects practical system-level conclusions, we compare complete system rankings under the SEEDA +Fluency setting. Table 13 reports the rankings induced by human judgments, ERRANT, SOME, and SURE.

The largest discrepancies for ERRANT occur for fluent rewrite-style systems. In both SEEDA-E and SEEDA-S, REF-F and GPT-3.5 are ranked first and second by humans but are placed 14th and 12th by ERRANT, respectively. SOME and SURE substantially reduce these rank displacements and recover the human top-two ordering.

SEEDA-E					SEEDA-S				
System	Human	ERRANT (Δ)	SOME (Δ)	SURE (Δ)	System	Human	ERRANT (Δ)	SOME (Δ)	SURE (Δ)
REF-F	1	14 (+13)	1 (+0)	1 (+0)	REF-F	1	14 (+13)	1 (+0)	1 (+0)
GPT-3.5	2	12 (+10)	2 (+0)	2 (+0)	GPT-3.5	2	12 (+10)	2 (+0)	2 (+0)
TransGEC	3	5 (+2)	4 (+1)	3 (+0)	T5	3	6 (+3)	3 (+0)	4 (+1)
T5	4	6 (+2)	3 (-1)	4 (+0)	TransGEC	4	5 (+1)	4 (+0)	3 (-1)
REF-M	5	13 (+8)	5 (+0)	5 (+0)	REF-M	5	13 (+8)	5 (+0)	5 (+0)
Riken-Tohoku	6	1 (-5)	8 (+2)	7 (+1)	BERT-fuse	6	3 (-3)	6 (+0)	6 (+0)
BERT-fuse	7	3 (-4)	6 (-1)	6 (-1)	Riken-Tohoku	7	1 (-6)	8 (+1)	7 (+0)
UEDIN-MS	8	2 (-6)	7 (-1)	8 (+0)	PIE	8	8 (+0)	9 (+1)	9 (+1)
PIE	9	8 (-1)	9 (+0)	9 (+0)	LM-Critic	9	9 (+0)	11 (+2)	11 (+2)
GECToR-BERT	10	7 (-3)	10 (+0)	10 (+0)	TemplateGEC	10	10 (+0)	12 (+2)	12 (+2)
LM-Critic	11	9 (-2)	11 (+0)	11 (+0)	GECToR-BERT	11	7 (-4)	10 (-1)	10 (-1)
GECToR-ens	12	4 (-8)	14 (+2)	14 (+2)	UEDIN-MS	12	2 (-10)	7 (-5)	8 (-4)
TemplateGEC	13	10 (-3)	12 (-1)	12 (-1)	GECToR-ens	13	4 (-9)	14 (+1)	14 (+1)
BART	14	11 (-3)	13 (-1)	13 (-1)	BART	14	11 (-3)	13 (-1)	13 (-1)

Table 13: System rankings under the SEEDA +Fluency setting. Human denotes the ranking induced by human judgments. Δ denotes the displacement from the human rank (metric rank minus human rank).

Metric	FCE Sys. (r/ρ)	Wiki Sys. (r/ρ)	Sys. Avg.	FCE Sent. (r/ρ)	Wiki Sent. (r/ρ)	Sent. Avg.
SOME	0.941/0.929	0.923/**0.833**	0.932/**0.881**	**0.653/0.643**	0.402/0.408	**0.528/0.526**
IMPARA	0.940/**0.952**	**0.954**/0.786	**0.947**/0.869	0.392/0.336	0.347/0.339	0.369/0.337
Scribendi	0.777/0.905	0.846/0.571	0.812/0.738	0.005/-0.039	0.210/0.180	0.108/0.070
SURE	**0.961**/0.929	0.694/0.595	0.827/0.762	0.575/0.553	**0.425/0.435**	0.500/0.494

Table 14: Zero-shot cross-domain evaluation on the FCE and Wiki domains of GMEG. System-level and sentence-level results are reported using Pearson's r and Spearman's ρ. SURE is evaluated without additional fine-tuning. The best score in each column and correlation measure is shown in bold.

Criteria	All Pairs		$\lvert\Delta h\rvert \geq 1.0$	
	SOME	SURE	SOME	SURE
Grammaticality	0.738	0.680	0.944	0.852
Fluency	0.698	0.669	0.911	0.871
Faithfulness	0.841	0.538	0.988	0.734

Table 15: Criteria-level pairwise accuracy on TMU-GFM. SOME is an in-domain model trained on TMU-GFM, while SURE is evaluated zero-shot. All pairs includes all candidate pairs with different human scores, and $|\Delta h| \geq 1.0$ keeps only confident pairs whose human score difference is at least 1.0. TMU-GFM's meaning-preservation axis is mapped to faithfulness for comparison.

C Additional Analyses

C.1 Cross-Domain Transfer on GMEG

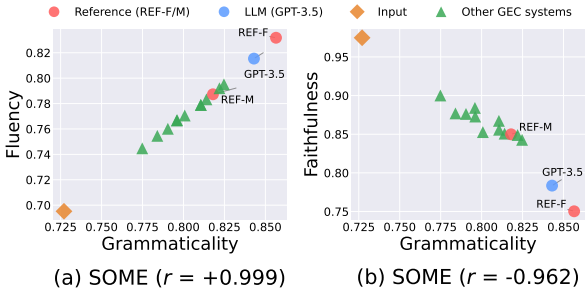
To assess cross-domain generalization, we evaluate SURE on the FCE and Wiki domains of GMEG (Napoles et al., 2019) without additional fine-tuning. We compare SURE with the reference-free baselines SOME, IMPARA, and Scribendi, and report Pearson's r and Spearman's ρ at both the system and sentence levels. As shown in Table 14, SURE shows competitive performance on FCE and stronger sentence-level results on Wiki, while its Wiki system-level correlations remain lower than those of SOME and IMPARA. These results suggest reasonable cross-domain transfer at the sentence level, while system-level calibration remains sensitive to domain shift.

C.2 Criteria-Level Transfer on TMU-GFM

Table 15 evaluates criteria-level transfer on TMU-GFM. SOME performs better across all axes on all pairs, which is expected given its in-domain axis-level supervision. However, when restricting the evaluation to confident pairs with human score differences of at least 1.0, SURE achieves reasonable zero-shot accuracy on grammaticality and fluency. Faithfulness remains the weakest axis, reflecting the difficulty of aligning our faithfulness criterion with TMU-GFM's meaning-preservation annotations.

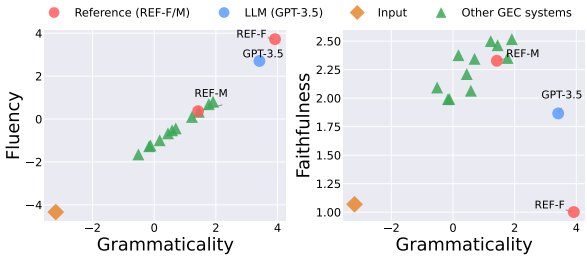
C.3 Additional Criteria-Pair Diagnostics

Figures 6 and 7 report the remaining criteria-pair correlations, using SOME's meaning-preservation score as the counterpart of faithfulness. Both SOME and SURE show a near-perfect positive correlation between grammaticality and fluency, indicating that the two criteria capture closely re-



(a) SOME (r = +0.999) (b) SOME (r = -0.962)

Figure 6: Additional criteria-pair correlations for SOME.



(a) SURE (r = +0.995) (b) SURE (r = +0.121)

Figure 7: Additional criteria-pair correlations for SURE.

lated aspects of correction quality. The distinction lies in faithfulness-related pairs: SOME shows a strong negative correlation between grammaticality and faithfulness, mirroring the faithfulness–fluency entanglement observed in the main analysis. In contrast, SURE keeps grammaticality and faithfulness much less correlated, indicating that faithfulness is not simply determined by either grammaticality or fluency. These results suggest that SURE provides a more separated criteria space for diagnosing system-level correction behavior.

D Prompt Templates

D.1 Correction Generation Prompts

Prompt for Minimal-Edit Correction

You are a careful English writing assistant.

Correct the following sentence:

Source:
{source}

Return only the corrected sentence on one line, with no commentary.

Prompt for Rewrite-Oriented Correction

You are a careful English writing assistant.

Rewrite the following sentence so that it sounds natural and fluent in English. You may restructure clauses, change word order, swap synonyms, or split/merge clauses as long as the original meaning is preserved. Do not add information that is not in the source. Aim for a noticeably different surface form from a minimal-edit correction.

Source:
{source}

Return only the rewritten sentence on one line, with no commentary.

D.2 GPT-4.1 Evaluation Prompts

GPT-4.1-E Evaluation Prompt

You are evaluating English grammatical error corrections.

Rank the {n} candidates by how well their **EDITS** (changes vs. the source) fix the source's errors.

Focus only on the quality of the edits:

- Reward edits that correctly fix grammatical, fluency, or usage errors.
- Penalize edits that introduce new errors, change the meaning, or make unnecessary changes.
- Do not reward a candidate simply for being fluent if its edits are not appropriate.
- If two candidates are similar, prefer the one that fixes more source errors with fewer unnecessary edits.

Source:
{source}

Candidates:
{candidate_block}

Output only the ranking from best to worst using candidate IDs.

GPT-4.1-S Evaluation Prompt

You are evaluating English sentences as grammatical corrections of a source.

Rank the {n} candidates by **OVERALL sentence quality**.

Focus on the final corrected sentence:

- Reward candidates that are grammatical, fluent, natural, and clear.
- Penalize candidates that contain grammar errors, awkward wording, or unnatural phrasing.
- Penalize candidates that distort or omit important meaning from the source.
- If two candidates are similar, prefer the one that is more natural and faithful.

Source:
{source}

Candidates:
{candidate_block}

Output only the ranking from best to worst using candidate IDs.

D.3 Pairwise GEC Annotation Prompt

Prompt for Pairwise GEC Annotation

You are an expert linguistic annotator for grammatical error correction (GEC). You compare two candidate corrections of the same source sentence on multiple axes. You always reply with a single JSON object that follows the requested schema exactly. Do not include commentary outside the JSON.

Source sentence:
`{source}`

Known errors in the source extracted by ERRANT against a minimal-edit reference; treat these as the ground-truth list of errors that should be addressed:
`{error_block}`

Candidate A:
`{cand_a}`

Candidate B:
`{cand_b}`

Your task
For each candidate, judge whether it RESOLVED each listed error.

- `"resolved"`: the error is corrected and the result is grammatical.
- `"missed"`: the error is still present or re-introduced.
- `"partial"`: attempted but incomplete, or introduces a new issue.

Then compare A vs B on three sentence-level axes:

- **Grammaticality**: which candidate has fewer grammar errors w.r.t. the source.
- **Faithfulness**: which candidate better preserves the source's meaning and intent.
- **Fluency**: which candidate reads as more natural, idiomatic English.

Finally, give an OVERALL preference for the better correction.

Important

- The amount of editing is not a criterion. Minimal edits and rewrites are equally valid; judge only on the criteria above.
- For each axis and overall, you must choose exactly one of `"A"` or `"B"`. Do not use `"tie"`. If the two candidates seem close on a criterion, still pick whichever is even marginally better.

Reply with a single JSON object using this exact schema:

```
{
  "span_resolution": {
    "A": {"e1": "resolved|missed|partial", ...},
    "B": {"e1": "resolved|missed|partial", ...}
  },
  "axes": {
    "grammaticality": "A|B",
    "faithfulness": "A|B",
    "fluency": "A|B"
  },
  "overall": "A|B",
  "rationale": "<one short sentence>"
}
```