

FECE: A Challenge Set for Evaluating Fine-Grained Translation Error Correction

Jie Ding, Xiangyu Duan

School of Computer Science and Technology, Soochow University, China
jding0508@stu.suda.edu.cn; xiangyuduan@suda.edu.cn

Abstract

Existing machine translation evaluation metrics assess translation quality only at the segment level and therefore cannot assess fine-grained translation quality. For translation error correction systems, this means that the quality of fine-grained corrections cannot be evaluated using existing metrics. To highlight this evaluation challenge, we propose the FECE dataset for evaluating fine-grained correction quality. Experiments on FECE show that existing metrics do not always reflect fine-grained correction performance. To enable the evaluation of fine-grained correction performance, we further propose $\text{COMET}_{\text{Fine}}$, a fine-grained evaluation metric. $\text{COMET}_{\text{Fine}}$ can effectively assess fine-grained correction quality while also revealing inconsistencies between existing metrics and $\text{COMET}_{\text{Fine}}$.

1 Introduction

Existing machine translation evaluation metrics primarily assess the quality of a complete segment. However, segment-level evaluation cannot distinguish corrections associated with human-annotated errors from changes made elsewhere in a translation, making it difficult to evaluate fine-grained translation quality independently.

This limitation is particularly important when evaluating translation error correction systems. A corrected translation may repair human-annotated errors in the original machine translation while also polishing or rewriting other parts of the segment. A segment-level score reflects both types of changes. Consequently, even when a corrected translation receives a higher score, it remains unclear whether the improvement results from correcting the annotated errors or from polishing and rewriting elsewhere. Existing segment-level metrics therefore cannot separate fine-grained correction quality from the overall quality of the complete translation.

To expose this evaluation challenge, we introduce the **Fine-Grained Error Correction Evaluation Set (FECE)**. Each FECE item contains two candidate translations generated from the same original machine translation. *hyp1* is instructed to correct all human-annotated errors while preserving the remaining content as closely as possible. In contrast, *hyp2* is instructed to retain all human-annotated errors while polishing content outside the annotated error locations. This controlled contrast allows us to test whether an evaluation metric reflects the quality of corrections to human-annotated errors or is influenced by polishing and rewriting elsewhere.

To enable the evaluation of fine-grained correction performance, we further propose $\text{COMET}_{\text{Fine}}$, a fine-grained evaluation metric. The metric first uses GPT-5.6 Sol (OpenAI, 2026) to align each human-annotated error with its corresponding correction content in *hyp1* or *hyp2*. The aligned correction content is then substituted for the corresponding human-annotated error in the original machine translation to obtain a reconstructed translation. Finally, the reconstructed translation is scored with COMET-22. By excluding content changes unrelated to the human-annotated errors, $\text{COMET}_{\text{Fine}}$ provides a more focused evaluation of fine-grained correction quality.

We use FECE to analyze the evaluation results of 29 WMT26 evaluation systems. The results show that existing metrics do not always reflect fine-grained correction performance. Across English–German, English–Japanese, and English–Chinese, WMT26 evaluation systems prefer *hyp1* in 51.2%, 39.2%, and 41.2% of cases on average, respectively. These results indicate that whole-segment scores can be influenced by polishing and rewriting elsewhere and therefore do not consistently reflect the quality of corrections to human-annotated errors. By contrast, $\text{COMET}_{\text{Fine}}$ prefers *hyp1* in 63.9%, 61.5%, and 68.4% of the pairs in the

three language directions, respectively, providing a clearer signal of fine-grained correction performance. Meanwhile, for many examples, WMT26 evaluation systems and COMET_{Fine} produce different preferences between *hyp1* and *hyp2*, further indicating that whole-segment evaluation does not always reflect fine-grained correction quality.

Our main contributions are threefold:

- We introduce FECE, a challenge set containing 3,000 candidate pairs across English–German, English–Japanese, and English–Chinese for evaluating fine-grained translation error correction quality.
- We systematically analyze the performance of 29 WMT26 evaluation systems on FECE and reveal the limitations of existing segment-level metrics in evaluating fine-grained corrections.
- We propose a fine-grained evaluation metric, COMET_{Fine}, which evaluates corrections corresponding to human-annotated errors through local alignment and reconstruction, and reveals inconsistencies between existing metrics and fine-grained evaluation results.

2 Related Work

Segment-level Machine Translation Evaluation and Error Localization Automatic machine translation (MT) metrics typically assign a segment-level quality score to a complete candidate translation. BLEU and chrF primarily assess translation quality based on word-level or character-level n-gram overlap between candidate and reference translations (Papineni et al., 2002; Popović, 2015). In contrast, BERTScore and YiSi leverage contextual representations from pretrained models to measure semantic similarity between candidate and reference translations (Zhang et al., 2020; Lo, 2019), while COMET learns to predict translation quality using pretrained multilingual representations and human quality judgments (Rei et al., 2020, 2022). In contrast to these automatic metrics, MQM and Error Span Annotation (ESA) provide human evidence about the locations and severities of translation errors (Freitag et al., 2021; Kocmi et al., 2024). xCOMET further combines segment-level scoring with automatic error localization (Guerreiro et al., 2024). These approaches improve the interpretability of MT evaluation, but

error detection and correction evaluation are distinct problems. The former determines whether and where an original translation is erroneous, whereas the latter must assess the quality of a candidate’s modification to a given human-annotated error. When a candidate also contains polishing or rewriting of other content, a segment-level score cannot separately reflect the changes resulting from corrections to human-annotated errors.

Translation error correction and minimal editing.

Automatic post-editing and translation error correction aim to improve MT output while preserving content that is already correct. Prior work has identified over-correction as a recurring problem and has used word-level quality-estimation signals to constrain generation and reduce unnecessary edits (Deoghare et al., 2025). The WMT25 shared task on automated translation evaluation also introduced quality-informed segment-level error correction, measuring overall improvement with the change in COMET score and editing efficiency with the Gain-to-Edit Ratio (Lavie et al., 2025). These methods consider both translation quality and the extent of editing, but the evaluation is still based on the complete corrected translation, making it difficult to distinguish the contribution of corrections to human-annotated errors from that of other content changes. Polishing content outside the human-annotated errors may increase the overall score, while other edits may offset the gains brought by error correction. Therefore, segment-level quality gains do not directly reflect the fine-grained correction quality associated with human-annotated errors.

Challenge sets for machine translation evaluation metrics.

Challenge sets are commonly used to examine the performance of machine translation systems or evaluation metrics in specific scenarios. Challenge sets for machine translation systems typically evaluate their ability to handle specific linguistic phenomena, including cross-lingual structural differences as well as lexical, grammatical, and discourse phenomena (Isabelle et al., 2017; Manakhimova et al., 2025). ACES applies this evaluation approach to automatic metrics by comparing better translations with translations containing specific accuracy errors, thereby analyzing metric sensitivity to different types of accuracy errors (Amrhein et al., 2022). Related WMT25 challenge sets further examine metric performance on low-quality and anomalous outputs (Knowles et al., 2025),

under-resourced Sub-Saharan African languages (Li et al., 2025), and gender bias (Filandrianos et al., 2025). These studies show that constructing dedicated challenge sets for specific translation phenomena, data conditions, or evaluation factors can help analyze the behavior and limitations of evaluation metrics across different settings.

Building on this line of work, FECE further focuses on the correspondence between edits in candidate translations and human-annotated errors. Given the same original machine translation, *hyp1* is instructed to correct all human-annotated errors while preserving the remaining content as much as possible, whereas *hyp2* retains these errors and only polishes content outside the annotated error regions. By comparing an evaluation metric’s preference between them, we can analyze which type of modification has a greater influence on its evaluation results. Based on this idea, COMET_{Fine} identifies the actual edits in a candidate translation that are associated with human-annotated errors and applies them to the original machine translation. It then scores the reconstructed translation, which retains only error-related edits, allowing the evaluation to focus more directly on the correction quality associated with human-annotated errors.

3 FECE Challenge Set

3.1 Data Construction

FECE is designed to examine whether automatic machine translation evaluation systems can reflect fine-grained translation error correction performance. The English–German data are derived from the WMT24 MQM data, whereas the English–Japanese and English–Chinese data are derived from the WMT24 ESA data (Freitag et al., 2024; Kocmi et al., 2024). These data provide source sentences, reference translations, original machine translation outputs, and corresponding human error annotations.

For each sample, we use the same source sentence, original machine translation, and human-annotated errors as input to DeepSeek-V4-Pro (DeepSeek-AI et al., 2026), which generates two candidate translations, *hyp1* and *hyp2*, under two complementary instructions:

- *hyp1* is instructed to correct all human-annotated errors in the original machine translation while preserving the remaining content as closely as possible.

- *hyp2* is instructed to retain all human-annotated errors in the original machine translation while polishing content outside the annotated error locations.

This contrastive design allows us to examine whether the scores produced by automatic evaluation systems reflect corrections to human-annotated errors or are instead influenced by polishing and rewriting elsewhere in the translation. If an evaluation system can accurately reflect fine-grained correction performance, it should, overall, show a preference for *hyp1*.

Direction	Source	Pairs	Candidates	Systems
En–De	WMT24 MQM	1,000	2,000	28
En–Ja	WMT24 ESA	1,000	2,000	28
En–Zh	WMT24 ESA	1,000	2,000	29
Total	—	3,000	6,000	29

Table 1: FECE data sources, scale, and WMT26 evaluation-system coverage.

3.2 Data Scale and Evaluation-System Coverage

FECE covers English–German, English–Japanese, and English–Chinese. Each language direction contains 1,000 candidate pairs, corresponding to 2,000 candidate translations. The complete challenge set therefore contains 3,000 candidate pairs and 6,000 candidate translations. Table 1 summarizes the data source, scale, and WMT26 evaluation-system coverage for each language direction.

The WMT26 automatic machine translation evaluation task provided scores from 29 distinct evaluation systems. The English–German and English–Japanese data were each scored by 28 systems, while the English–Chinese data were scored by 29 systems, yielding 85 system–language–direction evaluation results.

4 COMET_{Fine}

To evaluate fine-grained translation quality, we propose COMET_{Fine}. The method first identifies and aligns the edits in *hyp1* and *hyp2* that are related to human-annotated errors. It then applies these edits to the original machine translation to construct reconstructed translations that exclude unrelated polishing and rewriting, and finally scores the reconstructed translations using COMET-22.

Field	Text
Source	The storm left three people injured.
Original MT	<i>Der Sturm hinterließ drei Tote.</i>
Human annotation E1	<i>Tote</i>
hyp1	<i>Der Sturm hinterließ drei Verletzte.</i>
Alignment	original: Tote; replacement: Verletzte
hyp1_fine	<i>Der Sturm hinterließ drei Verletzte.</i>

Table 2: Example of candidate-specific alignment and reconstruction.

4.1 Local Alignment of Error-Related Edits

A complete candidate translation may contain both corrections related to the human-annotated errors and broader rewrites. For each candidate translation hyp_q ($q \in \{1, 2\}$), GPT-5.6 Sol receives the source sentence s , the original machine translation mt with human error annotations, and hyp_q . For each indexed error e_j , the model returns

$$a_{qj} = (o_{qj}, r_{qj}), \quad (1)$$

where `original` o_{qj} denotes the text span in mt associated with the j -th human-annotated error, and `replacement` r_{qj} denotes the replacement text aligned from the candidate translation. If the candidate does not modify the error, we set $r_{qj} = o_{qj}$; otherwise, r_{qj} records the text actually produced by the candidate at the corresponding location.

Table 2 shows an example of this alignment process. The German machine translation renders *injured* as *Tote* (“dead people”). In *hyp1*, the corresponding expression is *Verletzte* (“injured people”). The model therefore returns *Tote* as the original span and *Verletzte* as the replacement.

4.2 Reconstruction and Scoring

The aligned correction content is substituted for the corresponding human-annotated error span in the original machine translation:

$$\text{hyp}_q^{\text{fine}} = R\left(mt, \{o_{qj} \rightarrow r_{qj}\}_{j=1}^k\right). \quad (2)$$

We refer to the resulting reconstructed translations as *hyp1_fine* and *hyp2_fine*. Each reconstructed translation is based on the original machine translation mt and replaces only the `original` spans associated with human-annotated errors with their corresponding `replacement` spans from the candidate translation. Consequently, polishing and rewriting in the candidate translation that are unrelated to the human-annotated errors are not carried over into $\text{hyp}_q^{\text{fine}}$.

Let C denote COMET-22, which takes the source sentence s_i , a candidate translation, and the reference translation r_i as input. We define

$$\text{COMET}_{\text{Fine}}(\text{hyp}_{qi}) = C\left(s_i, \text{hyp}_{qi}^{\text{fine}}, r_i\right). \quad (3)$$

We rerun COMET-22 on both the complete candidate translations and the reconstructed translations in the same computational environment, thereby controlling for environment-dependent factors such as GPU numerical precision. To validate the reconstruction process, we randomly sample 200 pairs of reconstructed translations from each language direction for manual inspection, resulting in 600 inspected pairs in total. The reconstruction pass rates for English–German, English–Japanese, and English–Chinese are 98.5%, 99.0%, and 99.0%, respectively, with an overall pass rate of 98.8%.

5 Evaluation Setup

Our experiments examine three aspects: (1) the preferences of the WMT26 evaluation systems between the complete candidate translations *hyp1* and *hyp2*; (2) the preference expressed by $\text{COMET}_{\text{Fine}}$ between the two candidates after retaining only edits associated with human-annotated errors; and (3) the degree of agreement between the evaluation results of the WMT26 evaluation systems and those of $\text{COMET}_{\text{Fine}}$.

Pairwise Preference For each WMT26 evaluation system M , the score difference for the i -th pair of complete candidate translations is defined as

$$\Delta_{M,i} = M(\text{hyp1}_i) - M(\text{hyp2}_i). \quad (4)$$

Correspondingly, the score difference produced by $\text{COMET}_{\text{Fine}}$ is defined as

$$\Delta_{\text{Fine},i} = C\left(s_i, \text{hyp1}_i^{\text{fine}}, r_i\right) - C\left(s_i, \text{hyp2}_i^{\text{fine}}, r_i\right). \quad (5)$$

When *hyp1* receives a higher score than *hyp2*, the result is recorded as a preference for *hyp1*. When *hyp2* receives a higher score than *hyp1*, the result is recorded as a preference for *hyp2*. Equal scores are recorded as a tie. For each language direction, we separately calculate the proportions of *hyp1* preferences, *hyp2* preferences, and ties. If an evaluation system prefers *hyp1* in more pairs than it prefers *hyp2*, we consider that system to have an overall preference for *hyp1* in that language direction.

Agreement with COMET_{Fine} We compare the evaluation results of the WMT26 evaluation systems with those of COMET_{Fine} in terms of both pairwise preferences and score differences. Let $p_M(i)$ denote the preference assigned by evaluation system M to candidate pair i , and let $p_{\text{Fine}}(i)$ denote the preference assigned by COMET_{Fine} to the same candidate pair. Both preferences fall into one of three categories: a preference for *hyp1*, a preference for *hyp2*, or a tie.

If $p_M(i)$ and $p_{\text{Fine}}(i)$ are identical, the two evaluation results agree on that candidate pair. The three-way agreement rate is the proportion of all candidate pairs on which the evaluation results agree and is defined as

$$A(M) = \frac{1}{N} \sum_{i=1}^N \mathbf{1}[p_M(i) = p_{\text{Fine}}(i)]. \quad (6)$$

Here, N denotes the total number of candidate pairs, and $\mathbf{1}[\cdot]$ is the indicator function: it takes the value 1 when the condition inside the brackets is satisfied and 0 otherwise. All three categories—preferences for *hyp1*, preferences for *hyp2*, and ties—are included in the calculation. A tie assigned by both methods is counted as an agreement, whereas a tie assigned by only one method is counted as a disagreement.

The three-way agreement rate measures whether the two evaluation methods produce the same pairwise preferences. The Pearson correlation coefficient between Δ_M and Δ_{Fine} measures whether their score differences exhibit similar patterns of variation. When results are summarized within a language direction, we report the arithmetic mean of the system-specific agreement rates and Pearson correlation coefficients, with all available evaluation systems in that direction weighted equally. These values are descriptive macro-averages over evaluation systems. When results for an individual evaluation system are aggregated across language directions, preference and three-way agreement rates are averaged arithmetically. For Pearson correlation coefficients, we first apply Fisher’s z transformation, average the transformed values, and then transform the result back to the Pearson r scale.

6 Results

6.1 Preferences of WMT26 Evaluation Systems

Table 3 summarizes, by language direction, the mean preference rates of the WMT26 evaluation systems for *hyp1* and *hyp2*. It also reports the number of evaluation systems in each direction, the number of systems that prefer *hyp1* overall, and the mean agreement rate and Pearson correlation with COMET_{Fine}. For En–De, the 28 evaluation systems achieve an average *hyp1* preference rate of 51.2%, compared with a preference rate of 35.0% for *hyp2*. All 28 systems show an overall preference for *hyp1*. For En–Ja, the average *hyp1* preference rate across the 28 systems is 39.2%, lower than the *hyp2* preference rate of 50.0%, and only two systems show an overall preference for *hyp1*. For En–Zh, the 29 systems achieve an average *hyp1* preference rate of 41.2%, which is also lower than the *hyp2* preference rate of 47.1%. Only five systems show an overall preference for *hyp1* in this language direction.

The three language directions exhibit different evaluation patterns. The evaluation systems generally favor *hyp1* for En–De, whereas the average results favor *hyp2* for En–Ja and En–Zh. This suggests that, when complete candidate translations are evaluated, polishing and rewriting outside the human-annotated errors can substantially influence system scores. Consequently, *hyp1*, which is explicitly generated to correct the annotated errors, does not consistently receive higher scores across language directions.

At the individual-system level, the highest *hyp1* preference rates for En–De, En–Ja, and En–Zh are achieved by spBLEU (65.8%), BERTScore (54.2%), and BLEU (59.3%), respectively. Complete per-language results for all WMT26 evaluation systems are reported in Appendix D.

6.2 Evaluation Results of COMET_{Fine}

The *hyp1* preference rates of COMET_{Fine} are 63.9%, 61.5%, and 68.4% for En–De, En–Ja, and En–Zh, respectively. The corresponding *hyp2* preference rates are 28.5%, 28.9%, and 23.7% (Table 4 and Figure 1a). The mean fine-grained score difference is positive for all three language directions, indicating that COMET_{Fine} favors *hyp1* in each direction.

Compared with the average results of the WMT26 evaluation systems, the *hyp1* preference

Direction	Systems	<i>hyp1</i> pref.	<i>hyp2</i> pref.	<i>hyp1</i> -oriented	Agreement	Pearson r
En-De	28	51.2	35.0	28	53.3	0.444
En-Ja	28	39.2	50.0	2	46.1	0.214
En-Zh	29	41.2	47.1	5	46.3	0.309

Table 3: Average results of the WMT26 evaluation systems for each language direction. The *hyp1* preference rate, *hyp2* preference rate, and agreement rate are reported as percentages; the unreported proportion corresponds to ties. “*hyp1*-oriented” denotes the number of evaluation systems whose *hyp1* preference rate is higher than their *hyp2* preference rate in the corresponding language direction. Agreement and Pearson r measure the consistency between each evaluation system and COMET_{Fine} in terms of pairwise preferences and score differences, respectively.

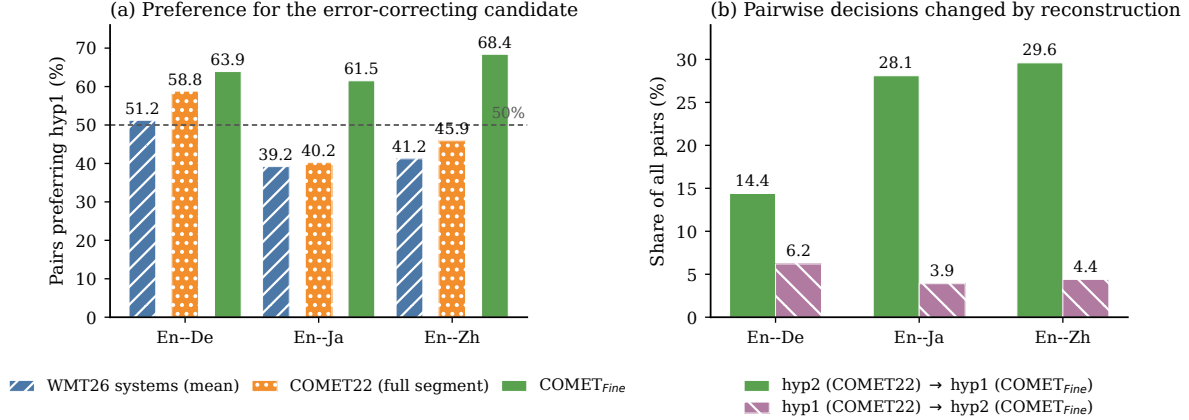


Figure 1: Main experimental results. (a) The average *hyp1* preference rate of the WMT26 evaluation systems, the *hyp1* preference rate of COMET-22 on the complete candidate translations, and the *hyp1* preference rate of COMET_{Fine}. (b) The proportions of candidate pairs whose COMET-22 preference changes from *hyp2* to *hyp1* or from *hyp1* to *hyp2* after the reconstructed translations are evaluated by COMET_{Fine}.

Direction	<i>hyp1</i> pref.	<i>hyp2</i> pref.	$\bar{\Delta}_{\text{Fine}}$
En-De	63.9	28.5	2.010
En-Ja	61.5	28.9	0.824
En-Zh	68.4	23.7	1.264

Table 4: Evaluation results of COMET_{Fine} for the three language directions. Preference rates are reported as percentages, and the unreported proportion corresponds to ties. A positive $\bar{\Delta}_{\text{Fine}}$ indicates that the average score of *hyp1_fine* is higher than that of *hyp2_fine*.

rates of COMET_{Fine} are higher by 12.7, 22.3, and 27.2 percentage points for En-De, En-Ja, and En-Zh, respectively. When the scope of evaluation is restricted to modifications associated with the human-annotated errors, *hyp1*, which is generated to correct these errors, exhibits a clearer advantage in all three language directions.

To further compare whole-segment and fine-grained evaluation, we use the same COMET-22 model to score both the complete candidate translations and the reconstructed translations. On the complete candidate translations, the *hyp1* preference rates of COMET-22 are 58.8%, 40.2%, and

45.9% for En-De, En-Ja, and En-Zh, respectively. On the reconstructed translations, the corresponding rates increase to 63.9%, 61.5%, and 68.4%, representing increases of 5.1, 21.3, and 22.5 percentage points. These changes demonstrate the influence of polishing and rewriting elsewhere in the candidate translations on whole-segment COMET-22 evaluation.

6.3 Differences between Whole-Segment and Fine-Grained Evaluation

Across all 3,000 candidate pairs, whole-segment COMET-22 prefers *hyp2* in 1,543 cases. Among these cases, 721 switch to a preference for *hyp1* under COMET_{Fine}. These 721 pairs account for 24.0% of all candidate pairs and 46.7% of the cases in which whole-segment COMET-22 prefers *hyp2*. Considering each language direction separately, 35.4%, 47.1%, and 54.8% of the En-De, En-Ja, and En-Zh pairs for which whole-segment COMET-22 prefers *hyp2* switch to a preference for *hyp1* under COMET_{Fine}, respectively.

In the opposite direction, 145 pairs switch from a

Direction	Agreement	Disagreement	Pearson r
En-De	72.3	27.7	0.797
En-Ja	58.6	41.4	0.588
En-Zh	58.2	41.8	0.642

Table 5: Agreement between whole-segment COMET-22 and COMET_{Fine} for each language direction. Agreement and disagreement rates are reported as percentages. Pearson r measures the correlation between their pairwise score differences.

preference for *hyp1* under whole-segment COMET-22 to a preference for *hyp2* under COMET_{Fine}. These cases account for 4.8% of all candidate pairs (Figure 1b; detailed results are provided in Appendix C). The two directions of change are markedly asymmetric: substantially more candidate pairs switch from *hyp2* to *hyp1* than from *hyp1* to *hyp2*.

These results suggest that part of the advantage obtained by *hyp2* in whole-segment evaluation may originate from polishing and rewriting outside the human-annotated errors. When these modifications are excluded from the reconstructed translations, the same COMET-22 model may instead prefer *hyp1*. Whole-segment evaluation results therefore cannot be directly interpreted as evaluations of the modifications associated with the human-annotated errors.

6.4 Agreement with COMET_{Fine}

The average three-way agreement rates between the WMT26 evaluation systems and COMET_{Fine} are 53.3%, 46.1%, and 46.3% for En-De, En-Ja, and En-Zh, respectively (Table 3). The corresponding average Pearson correlation coefficients are 0.444, 0.214, and 0.309. These results show that the WMT26 evaluation systems and COMET_{Fine} produce different preference judgments for many candidate pairs in FECE. Their score differences also do not exhibit strong parallel variation.

Among all evaluation systems, COMET-22 achieves the highest agreement rate and Pearson correlation coefficient in each of the three language directions, as shown in Table 5. However, full-segment COMET-22 and COMET_{Fine} still produce different preference decisions for 27.7%, 41.4%, and 41.8% of the candidate pairs in the three language directions, respectively. This disagreement is more pronounced for English-Japanese and English-Chinese, where it exceeds 40%.

Among the 28 evaluation systems for which

scores are available in all three language directions, BERTScore achieves the highest macro-average *hyp1* preference rate, at 58.4% (Appendix D). However, its macro-average three-way agreement rate with COMET_{Fine} is 54.0%, and its Fisher- z -aggregated Pearson correlation is 0.243. In comparison, COMET-22 has a lower macro-average *hyp1* preference rate of 48.3%, but a higher macro-average three-way agreement rate of 63.0% and a higher Fisher- z -aggregated Pearson correlation of 0.687. This comparison indicates that the *hyp1* preference rate and agreement with COMET_{Fine} capture two distinct aspects of evaluation: the former measures how frequently an evaluation system favors *hyp1*, whereas the latter measures whether its evaluation results are consistent with the fine-grained evaluation results.

6.5 Implications for Translation Error Correction Evaluation

The results on FECE expose an attribution problem in whole-segment evaluation. A change in the score of a corrected translation may result from the correction of human-annotated errors, but it may also result from polishing and rewriting elsewhere in the translation. An increase in the whole-segment score alone therefore cannot determine whether a translation error correction system has genuinely improved the content corresponding to the annotated errors. This issue is particularly evident for En-Ja and En-Zh: the average results of the WMT26 evaluation systems favor *hyp2*, whereas COMET_{Fine}, which evaluates only error-related modifications, favors *hyp1*. The difference between these evaluation results demonstrates that an improvement in overall translation quality does not necessarily imply an improvement in fine-grained correction quality.

Whole-segment and fine-grained evaluation capture different aspects of a translation error correction system. Whole-segment evaluation measures the overall quality of a corrected translation, whereas COMET_{Fine} focuses on the quality of modifications associated with human-annotated errors. When the two approaches produce different judgments, the change in overall translation quality and the effectiveness of correcting the target errors are not fully aligned. Translation error correction systems should therefore be evaluated by examining both aspects separately, rather than relying solely on whole-segment scores to assess correction performance.

7 Conclusion

We presented FECE, a 3,000-pair challenge set for evaluating the quality of fine-grained translation error correction. Across 29 WMT26 evaluation systems, whole-segment preferences do not consistently reflect whether the human-annotated errors were corrected. We also proposed COMET_{Fine}, which aligns annotated errors to candidate edits, reconstructs candidates from the original MT, and scores the resulting translations with COMET-22. COMET_{Fine} consistently favors the error-correcting candidate across all three language directions, preferring *hyp1* in 61.5%–68.4% of candidate pairs. Compared with COMET-22 evaluation on the complete segments, removing polishing and rewriting unrelated to the human-annotated errors causes 721 candidate pairs that were originally judged in favor of *hyp2* to instead favor *hyp1*. These results indicate that whole-segment evaluation can be influenced by modifications outside the annotated errors, whereas COMET_{Fine} more directly reflects the quality of corrections related to those errors. FECE and COMET_{Fine} therefore support the analysis of fine-grained correction quality at the levels of data construction and evaluation methodology, respectively.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62276179) and the Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

- Chantal Amrhein, Nikita Moghe, and Liane Guillou. 2022. [ACES: Translation accuracy challenge sets for evaluating machine translation metrics](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 479–513, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- DeepSeek-AI et al. 2026. [DeepSeek-V4: Towards highly efficient million-token context intelligence](#). *Computing Research Repository*, arXiv:2606.19348. Version 1.
- Sourabh Deoghare, Diptesh Kanojia, and Pushpak Bhattacharyya. 2025. [Giving the old a fresh spin: Quality estimation-assisted constrained decoding for automatic post-editing](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human*
- Language Technologies (Volume 2: Short Papers)*, pages 914–925, Albuquerque, New Mexico. Association for Computational Linguistics.
- George Filandrianos, Orfeas Menis Mastromichalakis, Wafaa Mohammed, Giuseppe Attanasio, and Chrysoula Zerva. 2025. [GAMBIT+: A challenge set for evaluating gender bias in machine translation quality estimation metrics](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 314–326, Suzhou, China. Association for Computational Linguistics.
- Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. [Experts, errors, and context: A large-scale study of human evaluation for machine translation](#). *Transactions of the Association for Computational Linguistics*, 9:1460–1474.
- Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. [Are LLMs breaking MT metrics? results of the WMT24 metrics shared task](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.
- Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. [xCOMET: Transparent machine translation evaluation through fine-grained error detection](#). *Transactions of the Association for Computational Linguistics*, 12:979–995.
- Pierre Isabelle, Colin Cherry, and George Foster. 2017. [A challenge set approach to evaluating machine translation](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2496, Copenhagen, Denmark. Association for Computational Linguistics.
- Rebecca Knowles, Samuel Larkin, and Chi-Kiu Lo. 2025. [MSLC25: Metric performance on low-quality machine translation, empty strings, and language variants](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 945–956, Suzhou, China. Association for Computational Linguistics.
- Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024. [Error span annotation: A balanced approach for human evaluation of machine translation](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453, Miami, Florida, USA. Association for Computational Linguistics.
- Alon Lavie, Greg Hanneman, Sweta Agrawal, Diptesh Kanojia, Chi-Kiu Lo, Vilém Zouhar, Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, Archchana Sindhuja, Jiayi Wang, David Ifeoluwa Adelani, Brian Thompson, Tom Kocmi,

- Markus Freitag, and Daniel Deutsch. 2025. [Findings of the WMT25 shared task on automated translation evaluation systems: Linguistic diversity is challenging and references still help](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 436–483, Suzhou, China. Association for Computational Linguistics.
- Senyu Li, Felermino Dario Mario Ali, Jiayi Wang, Rui Sousa-Silva, Henrique Lopes Cardoso, Pontus Stenertorp, Colin Cherry, and David Ifeoluwa Adelani. 2025. [Evaluating WMT 2025 metrics shared task submissions on the SSA-MTE African challenge set](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 913–919, Suzhou, China. Association for Computational Linguistics.
- Chi-kiu Lo. 2019. [YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources](#). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy. Association for Computational Linguistics.
- Shushen Manakhimova, Maria Kunilovskaya, Ekaterina Lapshinova-Koltunski, and Eleftherios Avramidis. 2025. [Fine-grained evaluation of English-Russian MT in 2025: Linguistic challenges mirroring human translator training](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 866–877, Suzhou, China. Association for Computational Linguistics.
- OpenAI. 2026. [GPT-5.6 system card](#).
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Maja Popović. 2015. [chrF: character n-gram F-score for automatic MT evaluation](#). In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.
- Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C. Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. [COMET-22: Unbabel-IST 2022 submission for the metrics shared task](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- 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.
- 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*.

A Candidate-Generation Prompts

For every FECE item, *hyp1* and *hyp2* were generated from the same English source sentence and the same original machine translation containing the human error tags. The following system prompts were used for the two candidates.

hyp1 Error-Correction Prompt

You are performing local error correction on a machine translation using the English source sentence and the human-labeled error spans.

Requirements:

1. Human-labeled errors are text enclosed by `<minor>...</minor>` or `<major>...</major>` tags.
2. The `minor` and `major` labels indicate the severity of the annotated error.
3. Correct every human-labeled error according to the English source sentence.
4. Only correct the human-labeled errors. Every labeled error must be genuinely corrected.
5. If changing only the exact labeled span is insufficient, you may extend the editing range as needed to fully correct the human-labeled error.
6. Do not polish, rewrite, reorganize, or optimize any unrelated content.
7. Output the complete corrected translation without error tags.

Return valid JSON only in the following format:

```
{
  "good": "complete corrected
translation"
}
```

hyp2 Fluency-Editing Prompt

You are performing local fluency editing on a machine translation while preserving all human-labeled errors and their original erroneous meanings.

Requirements:

1. Human-labeled errors are text enclosed by `<minor>...</minor>` or `<major>...</major>` tags.
2. The `minor` and `major` labels indicate the severity of the annotated error.
3. Preserve every human-labeled error and its original erroneous meaning.
4. Do not correct, weaken, remove, hide, reinterpret, or compensate for any labeled error.
5. Make multiple clearly visible fluency improvements outside the labeled error spans.
6. Do not rewrite the whole sentence or passage.
7. Output the complete translation without error tags.

Return valid JSON only in the following format:

```
{
  "bad": "complete locally
polished translation"
}
```

B GPT-Assisted Alignment of Error-Related Edits

The alignment of error-related edits was completed interactively with GPT-5.6 Sol through the Chat-GPT Work interface rather than through an API. Given an English source sentence, a machine translation with numbered human errors, and its corresponding hypothesis, the model identified the edit associated with each error and returned records in the following form:

```
{
  "error_ids": ["E1"],
  "original": "continuous span
from the MT",
  "replacement": "corresponding
continuous span from the HYP"
}
```

The `original` and `replacement` fields had to be continuous spans copied from the machine translation and hypothesis, respectively. The alignment covered the span directly related to the human-annotated error. The procedure recorded the observed edit without judging its correctness. An unchanged error therefore had identical `original` and `replacement` fields. A single edit involving multiple errors used a joint identifier list (e.g., `["E1", "E2"]`).

C Detailed Comparison Between Full-Segment COMET and COMET_{Fine}

Direction	Full-segment COMET-22			COMET _{Fine}			Preference shift		Agree	r
	<i>hyp1</i>	Tie	<i>hyp2</i>	<i>hyp1</i>	Tie	<i>hyp2</i>	<i>hyp2</i> $\rightarrow$ <i>hyp1</i>	<i>hyp1</i> $\rightarrow$ <i>hyp2</i>		
En-De	58.8	0.5	40.7	63.9	7.6	28.5	144 (14.4; 35.4)	62 (6.2)	72.3	0.797
En-Ja	40.2	0.2	59.6	61.5	9.6	28.9	281 (28.1; 47.1)	39 (3.9)	58.6	0.588
En-Zh	45.9	0.1	54.0	68.4	7.9	23.7	296 (29.6; 54.8)	44 (4.4)	58.2	0.642

Table 6: Comparison between full-segment COMET-22 and COMET_{Fine}. Preference rates and agreement are reported as percentages. *hyp2* $\rightarrow$ *hyp1* denotes a preference shift from *hyp2* under full-segment COMET-22 to *hyp1* under COMET_{Fine}. Results in this column follow the format “count (overall rate; conditional rate),” where the conditional rate is calculated over candidate pairs for which full-segment COMET-22 prefers *hyp2*. *hyp1* $\rightarrow$ *hyp2* denotes the reverse preference shift and follows the format “count (overall rate).” Pearson r is the Pearson correlation coefficient, computed over all candidate pairs, between the full-segment score differences Δ_M produced by the evaluation system and the reconstructed-translation score differences Δ_{Fine} produced by COMET_{Fine}.

Across all three directions, full-segment COMET-22 prefers *hyp2* in 1,543 of the 3,000 pairs. Of these, 721 switch to *hyp1* under COMET_{Fine}, accounting for 24.0% of all pairs and 46.7% of the pairs for which full-segment COMET-22 prefers *hyp2*. In the reverse direction, 145 pairs switch from *hyp1* to *hyp2*, accounting for 4.8% of all pairs.

D Results for All WMT26 Evaluation Systems

Tables 7, 8, and 9 report the per-system results for all WMT26 evaluation systems in the three language directions. The *hyp1* and *hyp2* preference rates are the percentages of candidate pairs for which an evaluation system assigns a higher score to *hyp1* and *hyp2*, respectively; cases in which the two candidates receive equal scores are counted as ties. The agreement rate is the percentage of candidate pairs for which the evaluation system and COMET_{Fine} make the same three-way decision among *hyp1*, tie, and *hyp2*. For each evaluation system, Pearson r denotes the Pearson correlation coefficient, computed across all candidate pairs in the corresponding language direction, between the system’s full-segment score differences Δ_M and the reconstructed-translation score differences Δ_{Fine} produced by COMET_{Fine}.

Evaluation system	<i>hyp1</i> pref.	Tie	<i>hyp2</i> pref.	Agreement	Pearson r
AEGIS	57.1	0.8	42.1	66.3	0.720
BERTScore	63.4	0.8	35.8	56.5	0.236
BLEU	63.3	4.0	32.7	54.6	0.156
COMET22	58.8	0.5	40.7	72.3	0.797
COMETKiwi22	56.5	1.5	42.0	58.2	0.598
Cohere CAT+ ensemble	39.2	34.2	26.6	43.6	0.518
FACET_869543	42.3	28.4	29.3	45.9	0.341
FACET_869546	42.7	28.0	29.3	45.8	0.343
Gemini-3.6-Flash	45.5	22.7	31.8	49.5	0.514
Gemma 4 - Reasoning Baseline	40.3	35.0	24.7	46.0	0.527
Gemma 4 Reference-based - Reasoning Baseline	42.0	30.2	27.8	46.7	0.537
MQM-LLM	44.1	25.4	30.5	48.4	0.503
MQM-LLM Confidence Aware	46.7	17.8	35.5	51.5	0.462
Qwen 3.6 - Reasoning Baseline	40.4	29.8	29.8	44.3	0.496
Qwen 3.6 Reference-based - Reasoning Baseline	39.4	33.7	26.9	43.6	0.473
Vertical_870257	50.2	10.6	39.2	53.1	0.368
Vertical_870554	53.9	8.6	37.5	52.4	0.396
YiSi-1	64.4	0.8	34.8	61.4	0.376
bytetop_pro10	61.4	0.1	38.5	63.9	0.643
bytetop_pro9	61.4	0.1	38.5	63.9	0.643
chrF	60.7	1.0	38.3	55.4	0.224
fluency2-gemini35-esa	43.9	15.5	40.6	47.4	0.052
gemba-poly	32.8	39.3	27.9	35.2	0.450
spBLEU	65.8	2.7	31.5	57.3	0.187
xCOMET XL Reference-based - Baseline	54.4	2.4	43.2	59.4	0.527
xCOMET XL Reference-free - Baseline	53.1	4.0	42.9	54.7	0.468
xCOMET XXL Reference-based - Baseline	57.6	2.7	39.7	58.8	0.506
xCOMET XXL Reference-free - Baseline	52.8	4.6	42.6	55.6	0.365

Table 7: Per-system results for all WMT26 evaluation systems on English–German. Preference and agreement rates are percentages.

Evaluation system	<i>hyp1</i> pref.	Tie	<i>hyp2</i> pref.	Agreement	Pearson <i>r</i>
AEGIS	39.3	0.2	60.5	51.9	0.342
BERTScore	54.2	0.2	45.6	52.9	0.238
BLEU	48.9	1.4	49.7	47.2	0.133
COMET22	40.2	0.2	59.6	58.6	0.588
COMETKiwi22	43.4	0.5	56.1	49.9	0.114
Cohere CAT+ ensemble	29.8	25.8	44.4	36.9	0.168
FACET_869543	30.4	25.2	44.4	40.1	0.105
FACET_869546	31.0	24.7	44.3	40.5	0.110
Gemini-3.6-Flash	32.3	22.1	45.6	38.5	0.214
Gemma 4 - Reasoning Baseline	32.7	26.5	40.8	40.7	0.244
Gemma 4 Reference-based - Reasoning Baseline	31.8	24.2	44.0	42.2	0.300
MQM-LLM	34.2	19.7	46.1	41.5	0.135
MQM-LLM Confidence Aware	37.0	13.9	49.1	42.6	0.145
Qwen 3.6 - Reasoning Baseline	31.3	26.8	41.9	38.0	0.173
Qwen 3.6 Reference-based - Reasoning Baseline	30.7	32.8	36.5	37.3	0.240
Vertical_870257	38.9	4.5	56.6	46.1	0.145
Vertical_870554	41.4	2.9	55.7	46.3	0.139
YiSi-1	49.6	0.3	50.1	51.7	0.255
bytetop_pro10	42.8	0.0	57.2	52.8	0.382
bytetop_pro9	42.8	0.0	57.2	52.8	0.382
chrF	46.6	0.8	52.6	47.8	0.189
fluency2-gemini35-esa	30.4	11.7	57.9	42.0	0.032
gemba-poly	29.7	33.6	36.7	36.7	0.116
spBLEU	50.4	1.5	48.1	47.9	0.160
xCOMET XL Reference-based - Baseline	45.2	0.3	54.5	54.2	0.303
xCOMET XL Reference-free - Baseline	40.8	0.4	58.8	49.0	0.177
xCOMET XXL Reference-based - Baseline	46.4	0.9	52.7	53.3	0.279
xCOMET XXL Reference-free - Baseline	44.8	2.2	53.0	51.1	0.183

Table 8: Per-system results for all WMT26 evaluation systems on English–Japanese. Preference and agreement rates are percentages.

Evaluation system	<i>hyp1</i> pref.	Tie	<i>hyp2</i> pref.	Agreement	Pearson <i>r</i>
AEGIS	45.3	0.1	54.6	54.9	0.514
BERTScore	57.6	0.1	42.3	52.7	0.255
BLEU	59.3	0.3	40.4	51.5	0.152
COMET22	45.9	0.1	54.0	58.2	0.642
COMETKiw22	41.0	0.6	58.4	48.7	0.417
Cohere CAT+ ensemble	29.4	30.1	40.5	38.1	0.281
FACET_869543	31.3	28.3	40.4	38.3	0.206
FACET_869546	31.2	28.2	40.6	38.1	0.209
Gemini-3.6-Flash	36.3	20.2	43.5	41.2	0.304
Gemma 4 - Reasoning Baseline	35.1	26.0	38.9	44.0	0.342
Gemma 4 Reference-based - Reasoning Baseline	33.8	24.8	41.4	42.7	0.355
MQM-LLM	33.9	21.2	44.9	41.1	0.254
MQM-LLM Confidence Aware	35.5	14.2	50.3	43.9	0.250
Qwen 3.6 - Reasoning Baseline	31.6	29.0	39.4	38.6	0.295
Qwen 3.6 Reference-based - Reasoning Baseline	33.0	32.4	34.6	39.7	0.296
Vertical_870257	38.3	4.3	57.4	45.2	0.294
Vertical_870554	41.7	2.9	55.4	47.5	0.278
YiSi-1	53.2	0.1	46.7	50.7	0.295
bytepop_pro10	49.0	0.0	51.0	53.2	0.506
bytepop_pro9	49.2	0.0	50.8	52.8	0.504
chrF	54.2	0.2	45.6	51.2	0.124
cuni-v14-regression	34.1	24.2	41.7	44.3	0.280
fluency2-gemini35-esa	30.6	14.8	54.6	40.8	0.123
gemba-poly	30.4	35.9	33.7	35.6	0.311
spBLEU	56.3	0.3	43.4	51.5	0.163
xCOMET XL Reference-based - Baseline	46.9	0.1	53.0	51.0	0.375
xCOMET XL Reference-free - Baseline	43.4	0.1	56.5	48.0	0.287
xCOMET XXL Reference-based - Baseline	44.3	0.2	55.5	49.2	0.350
xCOMET XXL Reference-free - Baseline	43.5	0.2	56.3	48.8	0.287

Table 9: Per-system results for all WMT26 evaluation systems on English–Chinese. Preference and agreement rates are percentages.