Simul-COMET: A Quality Metric for Simultaneous Interpretation in Distant Language Pair Considering Word Order Difference

Kosuke Doi[1,*], **Mana Makinae**[2], **Yusuke Sakai**[2], **Hidetaka Kamigaito**[2], **Taro Watanabe**[2]
[1]Seikei University, [2]Nara Institute of Science and Technology (NAIST)
kosuke-doi@st.seikei.ac.jp, {makinae.mana.mh2, sakai.yusuke.sr9, kamigaito.h, taro}@is.naist.jp

Abstract

In simultaneous interpretation (SI), interpreters perform real-time translation by segmenting the source speech into chunks and translating them in the order they appear. Since surface-matching metrics such as BLEU correlate poorly with human evaluations, translation quality is often evaluated using neural metrics that measure semantic similarity, such as COMET. However, while SI translation ideally exhibits high monotonicity, COMET tends to assign higher scores to offline translations with long-distance reordering, because it is trained on such offline translation data. To address this gap, we propose Simul-COMET, a variation of COMET adapted for SI evaluation specifically designed for monotonicity. We train Simul-COMET on the SI-style translation data, which was converted from the offline translation of the COMET training data by leveraging large language models. In English–Japanese translation experiments, we demonstrate that Simul-COMET assigns higher scores to SI-style translations than to offline ones. Moreover, Simul-COMET shows stronger alignment with evaluation scores provided by professional interpreters than the original COMET. Simul-COMET is available at https://github.com/kosuked/simul-comet.

1 Introduction

Simultaneous interpretation (SI) refers to the real-time translation of spoken content from a source language into a target language, requiring translation to begin before the source sentence ends. This time constraint poses a challenge of balancing translation quality and latency trade-off: accurate translation needs more source context at the cost of increased latency, while reducing latency entails translating from partial input, increasing the risk of errors. Human interpreters address this issue by segmenting a source sentence into a smaller unit, i.e., chunks, and translating them in order (Jones, 2002; Yagi, 2000; He et al., 2016), and simultaneous speech translation (SiST) models employ the same mechanism through decoding policies, which determines whether to generate a partial hypothesis or wait for more input (Liu et al., 2020; Papi et al., 2022a, 2023). This leads SI sentences to exhibit segment-level monotonicity as shown in Figure 1, where the offline reference (`off-ref`) shows reordering from the source (`src`) order, whereas the SI-style reference (`si-ref`) aligns with it. This monotonicity is an important aspect of SI quality evaluation, especially in language pairs with distinct syntactic structures (Doi et al., 2024a; Makinae et al., 2024a).

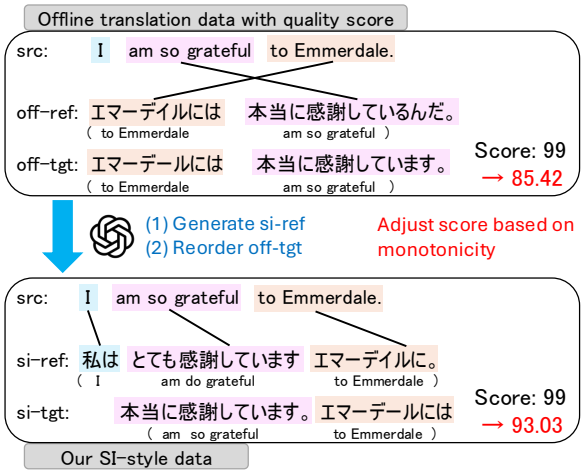

Figure 1: An example of offline and SI-style translation. The reference and target from offline data (off-ref and off-tgt, respectively) exhibit reordering of chunks compared to the source (src). Using an LLM, specifically GPT-4o-mini, we generate SI-style reference (si-ref) following Makinae et al. (2024a), and obtain SI-style target (si-tgt) by reordering off-tgt. In si-ref and si-tgt, the chunk order aligns with that of src.

The translation quality of SiST models is evaluated using popular metrics such as BLEU (Papineni

*This work was conducted while the author was affiliated with NAIST.

Findings of the Association for Computational Linguistics: ACL 2026, pages 42515–42533

et al., 2002), BLEURT (Sellam et al., 2020), and COMET (Rei et al., 2020), with little consideration for monotonicity.[1] One critical issue is that the references used in these metrics are often offline translations. Ko et al. (2023) and Zhao et al. (2024) address this by using SI data for model evaluation, and Doi et al. (2024a) showed that `si-ref` enable more accurate evaluation with metrics such as BLEU and BLEURT. However, COMET tends to assign higher scores to offline translations than to SI-style translations even when `si-ref` is used, because it is trained on offline translation data.

To address this gap, we propose Simul-COMET, a quality metric trained on SI-style data with consideration for monotonicity. COMET is a trainable neural metric that learns from source, human-scored target, and reference sentences. The default COMET model[2], `wmt22-comet-da`, is trained on Direct Assessment (DA) data, comprising human ratings from 0 to 100 assigned to outputs of offline machine translation (MT) models (Graham et al., 2013). However, collecting new data for SI-style targets with human evaluations is time-consuming and costly. Therefore, we leverage existing DA data and convert its offline target sentences (`off-tgt`) into SI-style translations (`si-tgt`) using a large language model (LLM), while preserving their content and errors (see Figure 1). We adjust the DA scores considering the monotonicity by penalizing them according to the degree of reordering between the source and target sentences. `si-ref` is also generated using an LLM from `src` (Sakai et al., 2024; Makinae et al., 2024a, 2026).

In this study, we develop Simul-COMET for SI between English and Japanese (en-ja), one of the most structurally different language pairs (SVO/head-initial vs. SOV/head-final), where translation monotonicity has a stronger impact on quality than in other pairs (Makinae et al., 2024a; Doi et al., 2024a). Due to this challenging nature, much prior work has focused primarily on this pair (Sakai et al., 2024; Zhao et al., 2024). Furthermore, in en-ja SI, many quality evaluation studies incorporate interpreters' perspectives (Tohyama and Matsubara, 2006; Okamura and Yamada, 2023), enabling the validation of automatic metrics against human judgments. Therefore, focusing on en-ja constitutes an essential first step for addressing monotonicity and ensuring reliable validation. We demonstrate that Simul-COMET assigns higher scores to `si-tgt` than to `off-tgt`. Moreover, Simul-COMET achieves a stronger correlation than default COMET with evaluations based on Multidimensional Quality Metrics (MQM)[3] conducted by two professional simultaneous interpreters.

2 Background and Related Work

2.1 Metrics for Translation Quality

Numerous automated metrics have been proposed for assessing the quality of MT outputs. Early metrics, e.g., BLEU (Papineni et al., 2002), rely on surface-level n-gram overlap. chrF++ (Popović, 2017) and RIBES (Isozaki et al., 2010) address the issue of BLEU for a more robust character-wise matching and distant reordering, respectively.

However, these surface-matching-based metrics often correlate poorly with human judgments (Sai et al., 2022), and neural-based metrics such as BLEURT (Sellam et al., 2020; Pu et al., 2021) and COMET (Rei et al., 2020, 2022) have addressed the issue by using contextual embeddings to measure semantic similarity. Despite the recent emergence of LLM-based metrics such as GEMBA (Kocmi and Federmann, 2023), pre-trained language model-based approaches like XCOMET (Guerreiro et al., 2023) and MetricX (Juraska et al., 2023) remained the top-performing metrics in the WMT24 Metrics Shared Task (Freitag et al., 2024).

2.2 Evaluation of SiST

Automatic evaluation of SiST typically relies on latency and quality metrics. Unlike latency, for which SiST-specific metrics have been developed (Ma et al., 2019a; Papi et al., 2022b; Kano et al., 2023; Papi et al., 2024; Makinae et al., 2025), quality is often assessed using the MT metrics described in Section 2.1. Macháček et al. (2023) reported that MT metrics correlate with human evaluations of SiST and recommended COMET as the preferred metric, while their study did not explore the metrics' ability to distinguish between offline and SI-style translations.

In contrast, Wein et al. (2024) argued that MT metrics are easily affected by SI-specific phenomena like segmentation and summarization. Word order differences were also found to influence their

[1] Some exceptions are Makinae et al. (2024a,b) and Sakai et al. (2024), which evaluate the monotonicity based on Spearman's rank correlation coefficient derived from source-target alignment.

[2] https://github.com/Unbabel/COMET

[3] https://themqm.org/

scores (Doi et al., 2024a). To improve alignment with human evaluations of SI, Wein et al. (2024) fine-tuned MetricX on SI data, but they did not incorporate monotonicity aspect, focusing only on adequacy and fluency. Makinae et al. (2024a) directly evaluated monotonicity by aligning source and target tokens with Awesome-Align (Dou and Neubig, 2021) and computing Spearman's correlation coefficient. Given that monotonicity has been actively studied in interpreting and translation studies, especially in the context of interpreter strategies and comparisons with other interpreting modes (Tohyama and Matsubara, 2006; Cai et al., 2020; Doi et al., 2024b), its incorporation into the automatic evaluation of SiST is an important challenge.

2.3 COMET

COMET is a translation quality evaluation metric built upon a pretrained multilingual encoder, such as XLM-R (Conneau et al., 2020). It takes the target, source, and reference as inputs, encoding them into a shared embedding space. These embeddings are combined to construct a feature, which is used as input to the feed-forward regressor (see Rei et al., 2020). The default COMET model, `wmt22-comet-da`, is trained on DA data to align with human preferences. DA is designed to evaluate machine-translated sentences based on adequacy and fluency, with human annotators, typically crowd workers, assigning scores on a 0 (poor) to 100 (perfect) scale (Graham et al., 2013).

XCOMET is a variant of COMET designed to identify error spans and compute an overall quality score. It uses MQM data for training, where translations are annotated by experts with span-level errors and their severity, i.e., 0 indicates no errors and larger values denote poorer translation quality. Other variants of COMET have also been proposed, including COMET-QE (Rei et al., 2021), which does not use references; models adapted to low-resource languages or those not supported by the default model (Sai B et al., 2023; Falcão et al., 2024); and models that incorporate context to perform document-level evaluation (Vernikos et al., 2022; Raunak et al., 2024).

2.4 Makinae et al.'s (2024a) SI Dataset Construction Methodology

Although several SI corpora containing human interpretations exist (Paulik and Waibel, 2009; Doi et al., 2021; Wang et al., 2021; Przybyl et al., 2022; Zhao et al., 2024), their creation is costly and time-consuming, motivating an approach that leverages LLMs to construct SI-style data automatically. Makinae et al. (2024a) prompted GPT-4o to segment source sentences into smaller units based on the salami technique (Jones, 2002) and translate them in order, enabling the generation of SI-like monotonic translations (henceforth Makinae et al.'s (2024a) Methodology). The salami technique is a strategy used by interpreters that segments sentences into smaller units with sufficient information for translation, allowing each segment to mirror source syntax and thus speed up translation. Makinae et al. (2024a) demonstrated that training SiST models on automatically generated SI data benefited grammatically distant language pairs, improving quality while reducing latency.

3 Simul-COMET

We propose Simul-COMET, a variant of COMET trained on SI-style DA (SI-DA) data instead of conventional offline data. Inspired by Makinae et al. (2024a), we generate SI-style versions of offline DA datasets using an LLM. We primarily focus on en-ja, a typologically distant language pair (e.g., word order: SVO vs. SOV), which is particularly challenging for SI (Sakai et al., 2024).

3.1 SI-DA Dataset Construction

As shown in Figure 1, COMET's DA dataset consists of three parts: source sentence (`src`), reference sentence (`off-ref`), and machine-translated target sentence (`off-tgt`). First, we generate SI-DA datasets by creating corresponding SI-style reference (`si-ref`) and target (`si-tgt`) sentences. Next, we apply a filtering step to ensure that the semantic coherence between `si-tgt` and `off-tgt`.

Settings We employ GPT-4o-mini (OpenAI et al., 2024), with all other settings following those of Makinae et al. (2024a)[4]. We use the en-ja subset of the DA data from the WMT 2020 Metrics Shared Task (Barrault et al., 2020)[5], which is divided into train, dev, and test splits containing 8,578, 500, and 500 sentences, respectively.

`si-ref` Following Makinae et al. (2024a), we leverage an LLM to segment each `src` into meaningful units based on the salami technique (Jones, 2002), and translate them in chunk order. We use the Makinae et al.'s (2024a) methodology.

[4] https://github.com/naist-nlp/SimulST

[5] https://github.com/Unbabel/COMET/tree/master/data

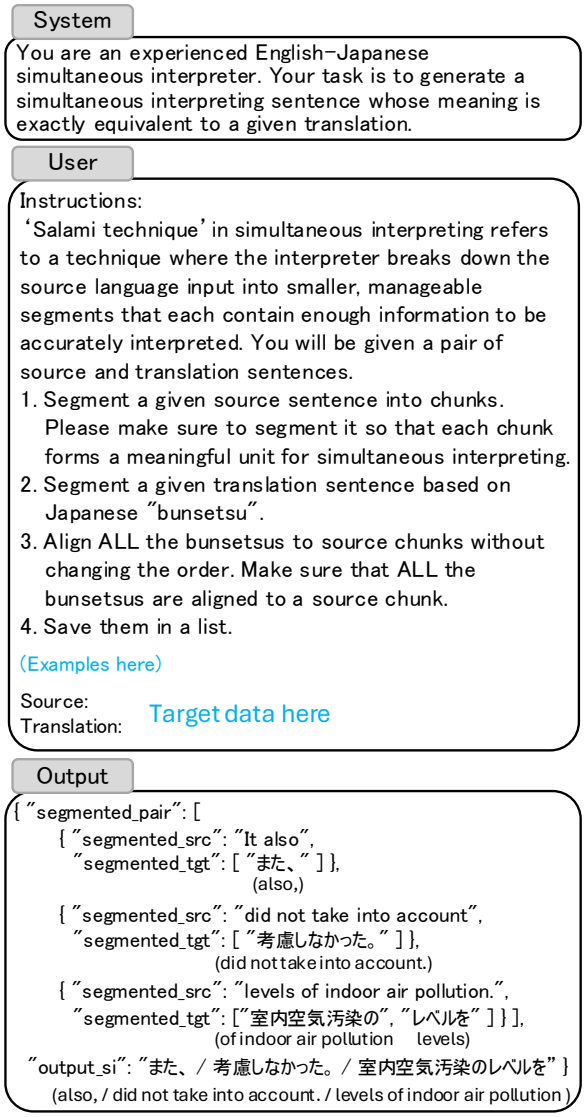

Figure 2: Prompt template for converting `off-tgt` in the DA data to `si-tgt`. We prompt GPT-4o-mini to segment the `src` and `off-tgt`, and align the segmented `off-tgt` phrases to the `src` chunks. The examples used are shown in Appendix A.

`si-tgt` Unlike `si-ref`, we need to create more controlled sentences around monotonicity and alignment for `si-tgt` in order to assign fine-grained DA scores. Therefore, we create `si-tgt` using the prompt shown in Figure 2. While Makinae et al.'s (2024a) methodology considers only `src`, we use both `src` and its corresponding `off-tgt`. First, we segment the `src` into syntactic chunk units, e.g., phrases, based on the salami technique, and the `off-tgt` into smaller units, i.e., *bunsetsus* which are Japanese phrase units[6]. Next, we align each unit of the `off-tgt` to one of the `src` chunks, even in cases where no clear correspondence exists due to mistranslation or other errors. We provide five-shot examples, each containing sentence-level errors, as well as under-translation and mistranslation at the chunk and phrase levels. Finally, we obtain the results in `JSON` format[7] to ensure that all steps are executed step by step in accordance with the instructions.

Threshold	Precision	Recall	F1
0.80	0.5121	**1.0000**	0.6773
0.85	0.5338	0.9961	0.6951
0.90	0.6250	0.8661	**0.7261**
0.95	**0.8493**	0.4882	0.6200

Table 1: Optimal threshold for the BERTScore filtering

Filtering While LLMs generally follow instructions well, they are not perfect and are known to occasionally fail at instruction following (Sakai et al., 2024, 2025). To ensure that the `si-tgt` preserves the same meaning as the `off-tgt`, without introducing any errors or hallucinations, we apply filtering based on BERTScore (Zhang et al., 2020), a token-level alignment-based similarity evaluation metric. We compute BERTScore between the `off-tgt` and `si-tgt`, and filter out the pair if any of the BERTScore precision, recall, or F1 scores fall below the threshold α. To determine an optimal threshold for filtering, we asked an annotator to assess the quality of all 500 sentences in the dev set, judging whether each `si-tgt` preserved the same meaning as the corresponding `off-tgt`. We computed precision, recall, and F1 scores for candidate thresholds $\{0.80, 0.85, 0.90, 0.95\}$, and found that $\alpha = 0.90$ yielded the highest F1 score, as shown in Table 1. Using this threshold, we finally obtained 6,078, 352, and 330 sentences for the train, dev, and test sets, respectively, removing 2,500 (29.1%), 148 (29.6%), and 170 (34.0%) sentences.

3.2 Adjusting DA Scores

The original DA scores, assigned by human annotators, are based on the accuracy of the translation.

[6] *Bunsetsus* are basic syntactic units of dependency in Japanese that consist of one or more content words followed by zero or more function words (Kawahara and Kurohashi, 2006). Preliminary experiments found that using the linguistic term bunsetsu slightly improved performance for Japanese compared to implicit segmentation based on the salami technique. This suggests that, when language-specific units such as *eojeol* in Korean are accessible, they may be more effective for other languages, while the salami technique can be applied elsewhere to support generalization.

[7] `https://platform.openai.com/docs/guides/structured-outputs`

To incorporate monotonicity with respect to the src, we adjust these scores accordingly. Specifically, we adjust the DA scores after min-max normalization[8], $\boldsymbol{DA} = DA_1, \ldots, DA_i, \ldots, DA_I$, where I denotes the number of sentence pairs, as defined by the following equation:

$$DA_i^m = DA_i - penalty_i^m \quad (1)$$

$$penalty_i^m = 0.25 \times (1 - MS_i), \quad (2)$$

where DA_i^m is the translation quality score for the i-th sentence, adjusted by its monotonicity score MS_i. Following Makinae et al. (2024a), MS_i is computed by aligning source and target tokens using Awesome-Align (Dou and Neubig, 2021), and then calculating Spearman's correlation coefficient[9]. MS_i equals 1 when the word order of the target sentence perfectly aligns with that of the source sentence. Accordingly, the greater the word order difference between the source and target, the lower the adjusted score DA_i^{m} becomes, compared to DA_i. We limit the maximum penalty applied to the score to 25%, inspired by the rule used for creating the DA Relative Ranks dataset (Ma et al., 2019b). This approach follows the common practice of subtractive penalties, such as the brevity penalty in BLEU (Papineni et al., 2002; Post, 2018), and enables modified DA scoring that penalizes based solely on monotonicity.

3.3 Model Training

We train Simul-COMET using the training data described in Section 3.1 and 3.2. Inspired by prior work on adapting COMET to unsupported or low-resource languages (Sai B et al., 2023; Falcão et al., 2024), we employ two training approaches: **(1) training models from scratch**, and **(2) fine-tuning the default COMET model**, i.e., `wmt22-comet-da` under three setting scenarios:

(a) Offline: We used the en-ja subset of the DA data without modification. Both the target and reference sentences are offline-based translations, i.e., `off-tgt` and `off-ref`.

(b) SI: We used our constructed SI-DA data. Both the target and reference sentences are SI-style translations, i.e., `si-tgt` and `si-ref`.

(c) Mixed: The target sentences include both offline-based and SI-style translations, while the reference sentences are all SI-style translations, i.e., `off-tgt`, `si-tgt`, and `si-ref`.

For each data type, we used two types of the DA scores: (i) **offline DA scores (da)**, which are directly assigned from the original offline dataset, and (ii) **monotonicity-adjusted SI-style DA scores (mono)**, which reflect monotonicity with respect to the source sentence structure.

Therefore, we trained models under six settings, corresponding to the combination of three data types and two DA score types: **(a-i)** ***off-da***, **(a-ii)** ***off-mono***, **(b-i)** ***si-da***, **(b-ii)** ***si-mono***, **(c-i)** ***mix-da***, and **(c-ii)** ***mix-mono***. For each setting, we applied both training approaches, i.e., training from scratch and fine-tuning from original COMET, resulting in a total of 12 models. Since we adjusted the scores of `off-tgt` and `si-tgt` in the same procedures (see Section 3.2) based on the degree of monotonicity with respect to the source sentence, the scores for `off-tgt` examples were lower than those for `si-tgt`. Therefore, we introduce no explicit signal for target sentence types, as the models can learn the input-score relationship from these data. In all settings, we trained the models for five epochs using mean squared error (MSE) loss. We followed the hyperparameter settings used in `wmt20-comet-da`[10]. For each setting, we selected the best checkpoint based on Kendall's rank correlation coefficient on dev set.

4 Experiments and Analysis

4.1 Analysis of SI-DA Dataset Quality

Settings We examined the quality of the SI-DA dataset from the perspectives of monotonicity and semantics. One purpose of constructing the SI-DA dataset is to create reference and target sentences monotonic to the source sentences. To examine whether the `si-ref` and `si-tgt` exhibit higher monotonicity than the `off-ref` and `off-tgt`, we calculated the monotonicity scores of these four types of sentences using the metric proposed by Makinae et al. (2024a) (see Section 2.2).

Moreover, in creating `si-tgt` from `off-tgt`, its content and errors should be preserved. We calculated BERTScore and COMET, using `si-tgt` as the hypothesis and `off-tgt` as the reference, to examine whether their meanings were equivalent. The quality of `si-ref` is assessed using COMET-

[8] We refer to the procedure for preparing DA scores for COMET training, as described in https://github.com/Unbabel/COMET/issues/131.

[9] We rescaled the scores from $[-1, 1]$ to $[0, 1]$.

[10] https://github.com/Unbabel/COMET/files/8502021/wmt20-comet-da-hyper-parameters.zip

	Reference		Target	
	off-ref	si-ref	off-tgt	si-tgt
Train	0.76 (0.19)	0.91 (0.11)	0.80 (0.18)	0.91 (0.10)
Dev	0.76 (0.20)	0.90 (0.12)	0.78 (0.18)	0.91 (0.11)
Test	0.77 (0.18)	0.91 (0.11)	0.81 (0.17)	0.90 (0.12)
Avg.	0.77	0.91	0.80	0.91

Table 2: Monotonicity scores for the reference and target sentences in the original DA and our SI-DA data (off-* and si-*, respectively). Values in parentheses indicate standard deviations (SDs).

	BERTScore			COMET
	Precision	Recall	F1	
Train	0.95 (0.03)	0.95 (0.03)	0.95 (0.03)	0.91 (0.06)
Dev	0.95 (0.03)	0.95 (0.03)	0.95 (0.03)	0.91 (0.06)
Test	0.96 (0.03)	0.95 (0.04)	0.95 (0.03)	0.91 (0.07)
Avg.	0.95	0.95	0.95	0.91

Table 3: Semantic similarity scores between off-tgt and si-tgt, calculated using the off-tgt as the reference. Values in parentheses indicate SDs.

Models	**Scratch**	**Fine-tuning**
wmt22-comet-da	−0.0366	–
off-da	−0.0284	−0.0373
off-mono	−0.0328	−0.0286
si-da	−0.0083	−0.0109
si-mono	−0.0105	−0.0068
mix-da	−0.0073	−0.0083
mix-mono	**+0.0087**	**+0.0030**

Table 4: Differences in Simul-COMET scores, calculated as the `si-tgt` score minus the `off-tgt` score. Positive values indicate that a model assigns higher score to `si-tgt` than to `off-tgt`, while negative values indicate the opposite. All differences between the `si-tgt` and `off-tgt` are statistically significant ($p < .05$).

QE (Rei et al., 2021), consistent with the validation protocol adopted by Makinae et al. (2024a).

Monotonicity Analysis Table 2 presents the monotonicity scores for the reference and target sentences in the original DA and our SI-DA data. The results show that the generated `si-ref` and `si-tgt` are more monotonic than translations in the original DA data. The average monotonicity scores for the en-ja subsets of MuST-C (Di Gangi et al., 2019) and Simul-MuST-C (Makinae et al., 2024a) are 0.7743 and 0.9073, respectively[11], which closely match the values in Table 2. These results suggest that our method successfully converted offline translations into SI-style by increasing their monotonicity.

Semantic Analysis Semantic similarity scores between `off-tgt` and `si-tgt` assessed using BERTScore and COMET are shown in Table 3. The scores were generally high, with all values exceeding 0.9. These results suggest that the converted `si-tgt` preserved the meanings and errors of the `off-tgt`, although human annotations show that some were not semantically identical despite high automatic scores.

The COMET-QE scores for `ref-si` were 0.8327, 0.8339, and 0.8371 for the train, dev, and test sets, respectively, indicating that the generated sentences are of relatively high and consistent quality across splits.

4.2 Performance of Simul-COMET

Settings To verify that Simul-COMET models assign higher scores to SI-style translations than to offline translations, we computed their scores on the test set using two types of target sentences: `off-tgt` and `si-tgt`. The `off-tgt` is taken from the original DA data, i.e., offline translation, while the `si-tgt` is SI-style translation generated by the method described in Section 3.1. We used `si-ref` as the reference data for both target sentences.

Simul-COMET Scores The difference in Simul-COMET score between `si-tgt` and `off-tgt` are shown in Table 4. The default COMET model, i.e., `wmt22-comet-da`, tends to assign higher scores to `off-tgt` than to `si-tgt`, which aligns with the results reported in Doi et al. (2024a). The same tendency was found in the proposed models: `off-tgt` received higher evaluations than `si-tgt` in all settings, except for the `mix-mono models`.

The performance of the `si-mono` models suggests that training Simul-COMET models using only SI-style target and reference sentences is insufficient. In our data, both `off-tgt` and `si-tgt` are penalized based on their monotonicity with respect to the source sentences. To enable models to assign scores that reflect the degree of monotonicity in a translation, it is essential to mix `off-tgt` and `si-tgt` examples so that models can learn the differences in their scores.

We further examined the effectiveness of using DA-SI data for training by calculating the win rate, i.e., the proportion of instances where `si-tgt` was

[11]We converted the scores reported in Makinae et al. (2024a), originally in the range of [-1, 1], to the [0, 1] range.

Models	Scratch	Fine-tuning
wmt22-comet-da	0.1429	–
off-da	0.1492	0.1134
off-mono	0.1371	0.1539
si-da	0.3388	0.1984
si-mono	0.2254	0.2520
mix-da	0.3443	0.3169
mix-mono	0.6748	0.6531

Table 5: The proportion of instances where si-tgt was assigned a higher score than off-tgt.

assigned a higher score than `off-tgt`. Table 5 indicates that the win rates of the `si-*` models are higher than those of the default and `off-*` models, suggesting that using solely SI-style data has a limited but positive effect on the model's tendency to assign higher scores to `si-tgt`. Only mixing the `off-tgt` and `si-tgt`, i.e., the `mix-da` setting, was also effective, whose win rates were higher than the default, `off-*`, and `si-*`. However, the results of `mix-mono`, which show a much higher win rate than the other settings, suggest that making Simul-COMET assign higher scores to `si-tgt` than to `off-tgt` requires training with mixed offline and SI data along with monotonicity-based score adjustment.

Score Distribution Analysis We observed that the score range of Simul-COMET differs across models; some models produced scores in a lower range than that of the default COMET (see Appendix D). Therefore, we examined the score distributions of the training data for COMET and its en-ja subset and found that the COMET data include more samples in higher-scoring range (mean=69.67, 68.68; SD=27.37, 20.09 for COMET and en-ja data, respectively). Furthermore, similar SDs were observed after monotonicity modification (mean=63.68, 66.33; SD=20.40, 20.24 for off-mono and si-mono, respectively), suggesting that the observed score range differences arise from the score distributions used during training.

We also observed differences in pseudo log-likelihood (PLL; Kauf and Ivanova, 2023) computed with XLM-R between `off-tgt` and `si-tgt`: PLL=-84.88, -91.83; length-normalized PLL=-2.60, -2.83, respectively. This difference suggests that XLM-R produces different sentence embeddings for `off-tgt` and `si-tgt`, which may influence the output score ranges of the trained Simul-COMET models. These findings suggest that differences in the target sentences themselves as well as differences in the quality-score distributions of the training data contribute to the observed differences in the Simul-COMET scores.

4.3 Human Evaluation

Settings To demonstrate that Simul-COMET models align with human evaluation of SI, we used the actual SI sentences produced by professional simultaneous interpreters. While Section 4.2 has shown that Simul-COMET captures monotonicity, this section examines its effectiveness in more realistic settings where human simultaneous interpretation reflects an interaction between fluency and monotonicity.

We extracted 243 en-ja SI segments from a press conference given by Aung San Suu Kyi at the Japan National Press Club[12]. We designed MQM-based guidelines[13] to independently assess SI quality in terms of accuracy, fluency, and monotonicity (see Appendix E), and asked two professional simultaneous interpreters, who did not perform the SI of this press conference, to evaluate its quality using the guidelines. Both interpreters who conducted the evaluation have over 20 years of SI experience.

Following the translation quality guidelines developed by the Japan Translation Federation (JTF)[14], E_i^c, the error score for error category c in the i-th sentence, was calculated as follows:

$$E_i^c = w_c \times (severity\ score) \tag{3}$$

where w_c is a weighting parameter representing the importance of error category c, selected from $\{0.5, 1.0, 1.5, 2.0\}$. The severity score is assigned as 100, 10, 1, and 0 for critical, major, minor, and none severities, respectively. The overall error score for the i-th sentence is computed as the sum of the error scores over all error categories. We report results using $w_c = 1.0$, i.e., the default in the JTF guidelines, for all error categories: accuracy, fluency, and monotonicity. We calculate correlation coefficient between this overall error score and Simul-COMET scores. We use SI sentences obtained in the same manner as the `si-ref` in the SI-DA dataset as references for computing Simul-COMET scores.

[12] https://www.jnpc.or.jp/archive/conferences/34352/report#

[13] The guidelines were reviewed by professional interpreters, and their feedback was incorporated into the final version.

[14] https://www.jtf.jp/tips/translation_quality_guidelines

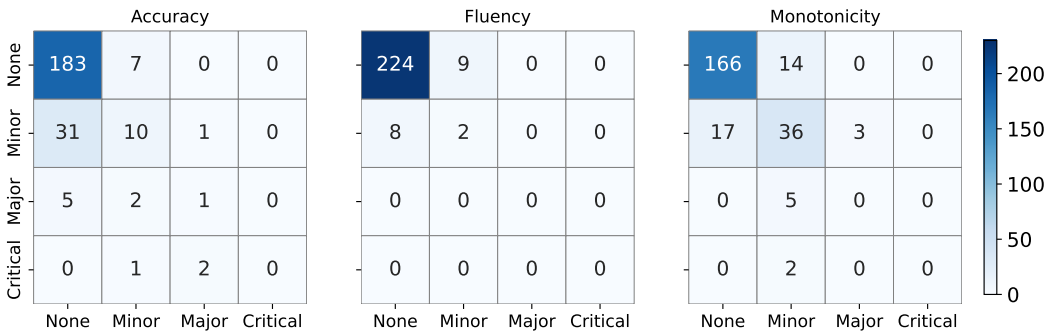

Figure 3: Human evaluation results conducted by two professional simultaneous interpreters

Categories	Pre.	Rec.	F1	QWK
Accuracy	0.7604	0.7984	0.7609	0.4488
Fluency	0.9333	0.9300	0.9316	0.1540
Monotonicity	0.8257	0.8313	0.8243	0.6260

Table 6: Inter-rater reliability based on multiclass precision (Pre.), recall (Rec.), and F1 scores and QWK

Metric	Rater A	Rater B	Avg. score
BLEU	0.0505	0.0917*	0.0597
wmt22-comet-da	0.1250	0.0874	0.1324*
mix-mono (scratch)	0.0817	0.0893	0.0903
mix-mono (fine-tuning)	**0.1418***	**0.1556***	**0.1719***

Table 7: Spearman's rank correlation coefficients between automatic metric scores and human evaluation scores. Avg. score indicates the correlation between automatic metric scores and the average of the scores provided by Rater A and Rater B. * indicates that the correlation coefficient was statistically significant ($p < .05$).

Inter-rater Reliability Figure 3 presents the evaluation results conducted by two professional simultaneous interpreters. The majority of SI sentences were assigned a 'none' rating for accuracy, fluency, and monotonicity (about 75% or more), suggesting that they were relatively high-quality SI sentences. For fluency and monotonicity, the two raters used each label at similar rates (fluency: about 95% and 5% for none and minor labels, respectively; monotonicity: about 75%, 23%, and 2% for none, minor, and major/critical labels, respectively). In contrast, for accuracy, one rater assigned the none label to only about 78% of the segments, whereas the other rater used it for about 90% of the segments, indicating differences in rating severity between the raters.

We calculated inter-rater reliability using multiclass precision, recall, and F1 scores, as well as Quadratic Weighted Kappa (QWK), with the results summarized in Table 6. The F1 scores were relatively high, exceeding 0.75 across the all evaluated categories. On the other hand, the QWK scores for accuracy and fluency were relatively low, corresponding to moderate and slight agreement according to the criteria of Landis and Koch (1977), respectively, suggesting a degree of inconsistency between the raters' assessments. The QWK score for monotonicity indicated substantial agreement, probably because it was evaluated by comparing the word order of the SI sentence with that of the source, while the other two categories largely rely on subjective interpretation by the raters.

Correlation with Automatic Metric Scores The correlation coefficients between human evaluation and Simul-COMET scores were computed using the `mix-mono` models, which demonstrated the ability to assign higher scores to `si-tgt` in the test set. Table 7 reports Spearman's rank correlation coefficient[15] between the overall error scores for the SI sentences and the Simul-COMET scores produced by the `mix-mono` models as well as BLEU and the default COMET model, `wmt22-comet-da`. Since the original human evaluation scores represent error values, i.e., lower values indicate better quality, we used their negative values to compute correlation coefficient. We computed the correlation with Rater A and Rater B individually, as well as with the average of their scores.

BLEU exhibits substantially weaker correlation with human judgments than the COMET-based models and its correlation coefficients are not statistically significant. Among the COMET-based models, the `mix-mono (fine-tuning)` model achieved the highest correlation with the averaged

[15]We observed similar trends using Kendall's rank correlation coefficient, the results of which are reported in Appendix F.

human scores ($\rho = 0.1719$, $p < 0.05$), followed by the `wmt22-comet-da` model ($\rho = 0.1324$, $p < 0.05$), while the `mix-mono (scratch)` model did not reach statistical significance ($\rho = 0.0903$, $p = 0.1606$).

We further conducted paired bootstrap resampling (Koehn, 2004) with 10,000 iterations, and observed that the `mix-mono (fine-tuning)` model outperformed the `wmt22-comet-da` model in 9,195 iterations, yielding a 90% confidence interval of [0.004, 0.160]. The `mix-mono (fine-tuning)` model also demonstrated a statistically significant correlation with the scores from both Rater A and B, whereas the other two models did not. These results suggest that fine-tuning approach improves the alignment between Simul-COMET predictions and human judgment in the evaluation of SI sentences. The poor performance of the `mix-mono (scratch)` model is likely due to the limited amount of training data.

As described in the setting section, the JTF guidelines allow users to assign different weights to error categories. We calculated correlation coefficients for all 64 possible combinations of error weights and ranked the three models. Table 8 reports the average ranking of each model, grouped by the relative weight of each error category. The results show that the model rankings remain unchanged across all weight configurations, with the `mix-mono (fine-tuning)` model ranked first, followed by the `wmt22-comet-da` model and the `mix-mono (scratch)` model. The consistent outperformance of the `mix-mono (fine-tuning)` model over the `wmt22-comet-da` model supports its effectiveness in capturing human preferences in SI evaluation.

Type	*N*	wmt22	scratch	fine-tuning
a > f > m	4	2.00	3.00	1.00
a > m > f	4	2.00	3.00	1.00
f > a > m	4	2.00	3.00	1.00
f > m > a	4	2.00	3.00	1.00
m > a > f	4	2.00	3.00	1.00
m > f > a	4	2.00	3.00	1.00
a = f > m	6	2.00	3.00	1.00
m > a = f	6	2.00	3.00	1.00
a = m > f	6	2.00	3.00	1.00
f > a = m	6	2.00	3.00	1.00
f = m > a	6	2.00	3.00	1.00
a > f = m	6	2.00	3.00	1.00
a = f = m	4	2.00	3.00	1.00

Table 8: Average ranking of the models for different error weights, grouped by the relative weight magnitude. a, f, and m denote accuracy, fluency, and monotonicity, respectively. wmt22, scratch, and fine-tuning represents `wmt22-comet-da`, `mix-mono (scratch)`, and `mix-mono (fine-tuning)` models, respectively. For example, a = f > m indicates that the weights assigned to accuracy and fluency are equal, and they are greater than the weight assigned to monotonicity, such as (a, f, m) = (1.5, 1.5, 1.0). N indicates the number of possible combinations for the type.

5 Conclusions

In this study, we proposed Simul-COMET, a metric for SI quality evaluation designed for monotonicity. In SI quality evaluation, translations with high monotonicity should be preferred, but COMET tends to assign higher scores to offline translations with reordering than to SI-style monotonic translations. Instead of collecting new data, we leveraged existing DA data by converting its offline sentences into SI-style sentences using LLMs, which were then used to train Simul-COMET models.

Our experiments in en-ja pair showed that Simul-COMET assigned higher scores to SI-style translations than to offline translations. We also demonstrated that Simul-COMET exhibited a stronger correlation with human evaluation scores by professional simultaneous interpreters than the original COMET. The results suggest that monotonicity-aware evaluation is a promising and necessary direction for SI evaluation. While this study focused on SI for en-ja, one of the most structurally divergent language pairs, future work will expand our method to a multilingual setting.

6 Limitations

Scalability of the other language pairs We focused on en-ja SI because this language pair exhibits significant structural differences, and translation monotonicity has a greater impact on translation quality compared to other language pairs (Makinae et al., 2024a; Doi et al., 2024a). Developing a monotonicity-aware evaluation method for one of the most challenging language pairs constitutes a crucial step toward the future extension to other language pairs. However, our proposed method for converting `off-tgt` into `si-tgt` relies on a characteristic of Japanese language that allows relatively flexible word order. Developing methods to construct SI-style data that reflect how human interpreters maintain monotonicity in target languages with strict word order constraints remains a challenge for future research.

Another challenge related to extending Simul-COMET to multiple languages is the difficulty of recruiting human annotators. Our approach of automatically creating training data using LLMs can reduce data construction costs, but automated evaluation metrics need to be assessed to determine whether they align with human judgments. Incorporating SI perspectives into evaluation requires experienced simultaneous interpreters, who are difficult to find and costly to engage. Furthermore, there is substantial research on en-ja SI quality that incorporates interpreters' perspectives (Tohyama and Matsubara, 2006; Okamura and Yamada, 2023), whereas interpreter-oriented evaluation criteria beyond the en-ja pair remain less well established. Due to limited available resources, this study conducted experiments only on en-ja, but future work should explore multiple languages.

Fluency of si-tgt Since our method reorders words in the `off-tgt` to create `si-tgt`, it may potentially produce unnatural Japanese sentences. We attempted to modify the prompt in Figure 2 by adding an instruction to naturally connect the Japanese segments, but this resulted in loss of alignment with the source word order. Sakai et al. (2025) observed that LLMs do not always follow instructions to generate sentences using specified words in a given order, with the best performance reaching only around 75%. They also argue that LLMs tend to maintain the naturalness of generated sentences. We hypothesize that adding an instruction to make the output more natural strengthened this tendency, leading LLMs to ignore the word order constraints. However, since humans can produce natural monotonic translations (Fukuda et al., 2024), generating similarly natural SI-style sentences with LLMs is a problem to be addressed in future research.

Impact of Using LLMs for Data Construction Although previous work has justified the use of LLMs (Cegin et al., 2023; Anikina et al., 2025), errors or biases introduced by an LLM may propagate into metrics. To ensure that LLM-generated SI sentences preserve the meaning of the original COMET DA data, we applied BERTScore-based filtering, and further assessed semantic preservation of the filtered sentences using BERTScore and COMET (see Sections 3.1 and 4.1, respectively). In addition, our mixed training setting combines LLM-generated SI sentences with the original COMET DA data, which reduces the impact of potential noise or bias introduced by the LLM.

Moreover, errors and biases are not unique to LLM-generated data because manually created data may also contain errors and stylistic variations due to individual differences among interpreters (Fukuda et al., 2024; Mizoguchi, 2009). Using LLMs also enables more consistent and cost-effective data creation, and Simul-COMET improved correlation with human judgments. However, our method creates `si-ref` using an LLM, specifically GPT-4o-mini, which makes the metric unsuitable for evaluating translation generated by the same LLM. Further analysis of LLM-generated data is required to fully clarify the limitations and advantages of LLM-based data construction.

Correlation with human evaluation scores Our proposed model showed only a weak correlation with the human evaluation scores provided by professional simultaneous interpreters, although its correlation coefficient was higher than that of `wmt22-comet-da`. While human evaluation was conducted along three dimensions, i.e., accuracy, fluency, and monotonicity, our model, `mix-mono (fine-tuning)` addressed only the monotonicity aspect. Since our offline-to-SI conversion method may degrade fluency in the target language, generating more fluent SI-style translations may improve the correlation with human judgment scores.

Furthermore SI-specific phenomena, e.g., summarization and omission, were not treated as errors in the human evaluation criteria for accuracy, whereas they may have been considered errors in the original DA scores. Simultaneous interpreters are known to apply interpretation strategies based on contextual understanding, and the correlation with human evaluation scores will improve further if quality scores can be designed to account for the use of such strategies.

7 Ethical Considerations

License of codes and dataset for training COMET models This study utilized the codes and dataset[16], which is governed by the Apache License 2.0. The license permits use, modification, and distribution of the software and associated data, including for commercial and research purposes, provided that proper attribution is given and the terms of the license are respected.

Ownership rights about outputs of the LLMs The SI-DA data was created using GPT-4o-mini

[16] https://github.com/Unbabel/COMET

and is therefore subject to OpenAI's license terms[17]. OpenAI assigns to us all rights, titles, and interests in and to the output.

Moderations The SI-DA data originates from a part of WMT20 DA data, whose source texts were extracted from online news websites (Barrault et al., 2020). Some of the texts include violent or sensitive content, e.g., wars and criminal cases. However, since these texts were originally published as news articles on websites, and the WMT20 dataset is widely used, we believe the data do not pose any critical ethical issues for research purposes.

Acknowledgments

We thank the anonymous reviewers and the area chair for their valuable comments. A part of this work has been supported by JSPS KAKENHI Grant Numbers JP21H05054, JP26K16092, 25K24369, and 26K21312, as well as JST SPRING Grant Number JPMJSP2140.

References

Tatiana Anikina, Jan Cegin, Jakub Simko, and Simon Ostermann. 2025. A rigorous evaluation of LLM data generation strategies for low-resource languages. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8282–8303, Suzhou, China. Association for Computational Linguistics.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, and 2 others. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Zhongxi Cai, Koichiro Ryu, and Shigeki Matsubara. 2020. What affects the word order of target language in simultaneous interpretation. In *Proceedings of 2020 International Conference on Asian Language Processing (IALP)*, pages 135–140.

Jan Cegin, Jakub Simko, and Peter Brusilovsky. 2023. ChatGPT to replace crowdsourcing of paraphrases for intent classification: Higher diversity and comparable model robustness. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1889–1905, Singapore. Association for Computational Linguistics.

[17] https://openai.com/policies/terms-of-use

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2012–2017, Minneapolis, Minnesota. Association for Computational Linguistics.

Kosuke Doi, Yuka Ko, Mana Makinae, Katsuhito Sudoh, and Satoshi Nakamura. 2024a. Word order in English-Japanese simultaneous interpretation: Analyses and evaluation using chunk-wise monotonic translation. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 254–264, Bangkok, Thailand (in-person and online). Association for Computational Linguistics.

Kosuke Doi, Katsuhito Sudoh, and Satoshi Nakamura. 2021. Large-scale English-Japanese simultaneous interpretation corpus: Construction and analyses with sentence-aligned data. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 226–235, Bangkok, Thailand (online). Association for Computational Linguistics.

Kosuke Doi, Katsuhito Sudoh, and Satoshi Nakamura. 2024b. Naist simultaneous interpretation corpus: Development and analyses of data from interpreters of different levels. *Journal of Natural Language Processing*, 31(3):868–893.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online. Association for Computational Linguistics.

Júlia Falcão, Claudia Borg, Nora Aranberri, and Kurt Abela. 2024. COMET for low-resource machine translation evaluation: A case study of English-Maltese and Spanish-Basque. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3553–3565, Torino, Italia. ELRA and ICCL.

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.

Ryo Fukuda, Kosuke Doi, Katsuhito Sudoh, and Satoshi Nakamura. 2024. 原発話に忠実な英日同時機械翻訳の実現に向けた順送り訳評価データ作成 [Creation of Evaluation Data for Monotonic Translation toward the Realization of Simultaneous English-Japanese Machine Translation Faithful to the Source Speech]. In *Proceedings of the 259th meeting of Special Interest Group of Natural Language Processing (IPSJ-SIGNL), 2024-NL-259(14)*, pages 1–6. (in Japanese).

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2023. xcomet: Transparent machine translation evaluation through fine-grained error detection. *arXiv*, arXiv:2310.10482.

He He, Jordan Boyd-Graber, and Hal Daumé III. 2016. Interpretese vs. translationese: The uniqueness of human strategies in simultaneous interpretation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 971–976, San Diego, California. Association for Computational Linguistics.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 944–952, Cambridge, MA. Association for Computational Linguistics.

Roderick Jones. 2002. *Conference Interpreting Explained Second Edition*. Routledge, New York.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics.

Yasumasa Kano, Katsuhito Sudoh, and Satoshi Nakamura. 2023. Average token delay: A latency metric for simultaneous translation. In *Interspeech 2023*, pages 4469–4473.

Carina Kauf and Anna A. Ivanova. 2023. A better way to do masked language model scoring. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 925–935, Toronto, Canada. Association for Computational Linguistics.

Daisuke Kawahara and Sadao Kurohashi. 2006. A fully-lexicalized probabilistic model for Japanese syntactic and case structure analysis. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 176–183, New York City, USA. Association for Computational Linguistics.

Yuka Ko, Ryo Fukuda, Yuta Nishikawa, Yasumasa Kano, Katsuhito Sudoh, and Satoshi Nakamura. 2023. Tagged end-to-end simultaneous speech translation training using simultaneous interpretation data. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 363–375, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Danni Liu, Gerasimos Spanakis, and Jan Niehues. 2020. Low-latency sequence-to-sequence speech recognition and translation by partial hypothesis selection. In *Interspeech 2020*, pages 3620–3624.

Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, Hua Wu, and Haifeng Wang. 2019a. STACL: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3025–3036, Florence, Italy. Association for Computational Linguistics.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019b. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

Dominik Macháček, Ondřej Bojar, and Raj Dabre. 2023. MT metrics correlate with human ratings of simultaneous speech translation. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 169–179, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Mana Makinae, Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2024a. Simul-MuST-C: Simultaneous multilingual speech translation corpus using large language model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22185–22205, Miami, Florida, USA. Association for Computational Linguistics.

Mana Makinae, Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2025. Investigating omission as a latency reduction strategy in simultaneous speech translation. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 2238–2258, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Mana Makinae, Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2026. Rethinking evaluation in simultaneous speech translation: A case for monotonic test sets. *Computational Linguistics*, pages 1–30.

Mana Makinae, Katsuhito Sudoh, Masaru Yamada, and Satoshi Nakamura. 2024b. An automatic quality metric for evaluating simultaneous interpretation. *arXiv*, arXiv:2407.06650.

Ryoko Mizoguchi. 2009. Study on the Interpreter's Role –Six Styles and Two Functions. *Interpreting and Translation Studies*, 9:71–86. (in Japanese).

Hajime Morita, Daisuke Kawahara, and Sadao Kurohashi. 2015. Morphological analysis for unsegmented languages using recurrent neural network language model. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2292–2297, Lisbon, Portugal. Association for Computational Linguistics.

Yuki Okamura and Masaru Yamada. 2023. 「順送り訳」の規範と模範 同時通訳を模範とした教育論の試論 [Norms and Canon of Progressive Translation - An Exploratory Study on Educational Theories Using Simultaneous Interpretation as a Canon]. In Hiroyuki Ishizuka, editor, *Word Order in English-Japanese Interpreting and Translation: The History, Theory and Practice of Progressive Translation*, pages 217–250. Hitsuji Syobo. (in Japanese).

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Sara Papi, Marco Gaido, Matteo Negri, and Luisa Bentivogli. 2024. StreamAtt: Direct streaming speech-to-text translation with attention-based audio history selection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3692–3707, Bangkok, Thailand. Association for Computational Linguistics.

Sara Papi, Marco Gaido, Matteo Negri, and Marco Turchi. 2022a. Does simultaneous speech translation need simultaneous models? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 141–153, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sara Papi, Marco Gaido, Matteo Negri, and Marco Turchi. 2022b. Over-generation cannot be rewarded: Length-adaptive average lagging for simultaneous speech translation. In *Proceedings of the Third Workshop on Automatic Simultaneous Translation*, pages 12–17, Online. Association for Computational Linguistics.

Sara Papi, Matteo Negri, and Marco Turchi. 2023. Attention as a guide for simultaneous speech translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13340–13356, Toronto, Canada. Association for Computational Linguistics.

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

Matthias Paulik and Alex Waibel. 2009. Automatic translation from parallel speech: Simultaneous interpretation as mt training data. In *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*, pages 496–501.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Heike Przybyl, Ekaterina Lapshinova-Koltunski, Katrin Menzel, Stefan Fischer, and Elke Teich. 2022. EPIC UdS - creation and applications of a simultaneous interpreting corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1193–1200, Marseille, France. European Language Resources Association.

Amy Pu, Hyung Won Chung, Ankur Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning

compact metrics for MT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 751–762, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vikas Raunak, Tom Kocmi, and Matt Post. 2024. SLIDE: Reference-free evaluation for machine translation using a sliding document window. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 205–211, Mexico City, Mexico. 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, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. Are references really needed? unbabel-IST 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online. 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.

Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2022. A survey of evaluation metrics used for nlg systems. *ACM Computing Surveys*, 55(2):1–39.

Ananya Sai B, Tanay Dixit, Vignesh Nagarajan, Anoop Kunchukuttan, Pratyush Kumar, Mitesh M. Khapra, and Raj Dabre. 2023. IndicMT eval: A dataset to meta-evaluate machine translation metrics for Indian languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14210–14228, Toronto, Canada. Association for Computational Linguistics.

Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2025. Revisiting compositional generalization capability of large language models considering instruction following ability. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31219–31238, Vienna, Austria. Association for Computational Linguistics.

Yusuke Sakai, Mana Makinae, Hidetaka Kamigaito, and Taro Watanabe. 2024. Simultaneous interpretation corpus construction by large language models in distant language pair. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22375–22398, Miami, Florida, USA. 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.

Hitomi Tohyama and Shigeki Matsubara. 2006. Collection of simultaneous interpreting patterns by using bilingual spoken monologue corpus. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Giorgos Vernikos, Brian Thompson, Prashant Mathur, and Marcello Federico. 2022. Embarrassingly easy document-level MT metrics: How to convert any pretrained metric into a document-level metric. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 118–128, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online. Association for Computational Linguistics.

Shira Wein, Te I, Colin Cherry, Juraj Juraska, Dirk Padfield, and Wolfgang Macherey. 2024. Barriers to effective evaluation of simultaneous interpretation. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 209–219, St. Julian's, Malta. Association for Computational Linguistics.

Sane Yagi. 2000. Studying style in simultaneous interpretation. *Meta*, 45(3):520–547.

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*.

Jinming Zhao, Yuka Ko, Kosuke Doi, Katsuhito Fukuda, Ryo Sudoh, and Satoshi Nakamura. 2024. NAIST-SIC-aligned: An aligned English-Japanese simultaneous interpretation corpus. In *Proceedings of the*

2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 12046–12052, Torino, Italia. ELRA and ICCL.

A The Example Sentences Used in the Prompt for Converting Offline Translations to SI-style Translations

In the prompt for converting `off-tgt` to `si-tgt`, as shown in Figure 2, we provided an LLM with the following five examples. // represents segment boundaries and / represents boundaries for Japanese *bunsetsus*. Note that the translation in the third example contains corresponding words, but the Japanese sentence is unnatural and does not convey coherent meaning.

source: I'm an economist and we conduct so-called experiments to address this issue.
translation: 私は経済学者で、解決するために実験をします。
segmented:
I 私は
'm an economist 経済学者で、
and we conduct so-called experiments 実験を / します。
to address this issue. 解決するために
output:
私は // 経済学者で、 // 実験をします。 // 解決するために [I // 'm an economist // conduct experiments // to address]

source: Thousands of people in the secretive nation were quarantined, but restrictions had recently eased.
translation: 秘密主義国家のたくさんの人が隔離されましたが、制限は最近緩和されました。
segmented:
Thousands of people たくさんの / 人が
in the secretive nation 秘密主義 / 国家の
were quarantined, 隔離 / されましたが、
but restrictions had recently eased. 制限は / 最近 / 緩和 / されました。
output:
たくさんの人が // 秘密主義国家の // 隔離されましたが、 // 制限は最近緩和されました。 [Many people // in the secretive nation // were quarantined, but // restrictions had recently eased]

source: I'm very disappointed and upset as we're packed and ready to go.
translation: 私は非常に失望していると私たちはパックされ、行く準備ができているように動揺。
segmented:
I'm very disappointed 私は / 非常に / 失望 / している
and upset と / 動揺。
as we're packed 私たちは / パックされ、
and ready to go. 行く / 準備が / できて / いるように
output:
私は非常に失望している // と動揺。 // 私たちはパックされ、 // 行く準備ができているように [I'm very disappointed // and upset // we're packed // ready to go]

source: Kanye West issues public apology to wife Kim Kardashian after marriage revelations.
translation: カニエ・ウェストは、妻キム・カーダシアンに謝罪する。
segmented:
Kanye West カニエ・ウェストは、
issues public apology 謝罪する。
to wife Kim Kardashian 妻 / キム・カーダシアンに
after marriage revelations
output:
カニエ・ウェストは // 謝罪する。 // 妻キム・カーダシアンに // [Kanye West issues public apology to wife Kim Kardashian]

source: To Kim I want to say I know I hurt you.
translation: キムには、私が何かをしたのを知っていると、私は言いたいです。
segmented:
To Kim キムには、
I want to say 私は / 言いたいです。
I know 知っていると、
I hurt you. 私が / 何かを / したのを
output:
キムには、 // 私は言いたいです。 // 知っていると // 私が何かをしたのを [To Kim // I want to say // I know // I did something]

B Comparison of Methods for Generating SI-style References

As an alternative of the method proposed by Makinae et al. (2024a), we attempted to generate `si-ref` using a similar approach to that used for creating `si-tgt`, i.e., segmenting `src` and `off-ref` into

Method	Human	BERTScore F1	Monotonicity	Fluency
Makinae et al. (2024a)	**35**	**0.6911**	0.8776	**4.3600**
Alignment	15	0.6814	**0.8863**	3.9800

Table 9: Comparison of methods for generating `si-ref`

Query	What is a bunsetsu in Japanese language?
Response	A bunsetsu can vary in length and complexity and is essential for understanding the structure of Japanese sentences. For example, a bunsetsu may consist of a noun and its modifier, a verb, or even a complete subject-predicate structure. In practice, identifying bunsetsu helps in parsing sentences for better comprehension and translation.

Table 10: The prompt for asking about Japanese `bunsetsu` and the response from GPT-4o-mini

chunks and aligning them. To compare the two methods, we used the first 50 sentences from the dev set, shuffling them so that the method used to generate each one was not identifiable. Then, the first author selected the sentence that was better as an SI sentence. The human judgment results along with automatic metric scores are shown in Table 9. The `si-ref` generated by the method of Makinae et al. (2024a) was more preferred by the human rater and received higher scores in automatic evaluation metrics (except for monotonicity), so we adopted their method in our experiments.

C Does an LLM Understand Japanese *Bunsetsu*?

Our preliminary analysis showed that LLMs possess knowledge of Japanese *bunsetsu* structures. Table 10 shows an example prompt and a response from the GPT-4o-mini, i.e., the LLM used in our study.

This suggests that LLMs have the ability to segment Japanese sentences into `bunsetsus`, but Japanese syntactic parsers that can perform such segmentation already exist, e.g., GinZA[18] and KNP (Kawahara and Kurohashi, 2006). In our preliminary experiments, we segmented Japanese sentences into *bunsetsus* using KNP, which takes morphological analysis results from Juman++ (Morita et al., 2015), finding that parsing failed for about 6.5% of the sentences in the dev set, e.g., in cases where an MT output contained special character types. We use *bunsetsus* in the process of converting `off-tgt` to `si-tgt`; however, removing special characters that causes parsing errors would conflict with our goal of preserving the errors and meanings of the `off-tgt`.

[18] https://github.com/megagonlabs/ginza

Models	Scratch		Fine-tuning	
	off-tgt	si-tgt	off-tgt	si-tgt
wmt22-comet-da	**0.8593**	0.8227	–	–
off-da	**0.8278**	0.7994	**0.6303**	0.5930
off-mono	**0.6356**	0.6028	**0.6314**	0.6028
si-da	**0.6548**	0.6465	**0.6428**	0.6319
si-mono	**0.6732**	0.6627	**0.6423**	0.6355
mix-da	**0.8440**	0.8367	**0.6337**	0.6254
mix-mono	0.7027	**0.7114**	0.6141	**0.6171**

Table 11: Simul-COMET scores on the test set. The references are ref-si in all settings. All differences between the `off-tgt` and `si-tgt` are statistically significant ($p < .05$). For each model setting, the higher score between off-tgt and si-tgt is shown in bold.

Therefore, we used an LLM, specifically GPT-4o-mini, for *bunsetsu* segmentation; however, its agreement with the KNP-based segmentation results on the dev set was not high, at 0.56. We analyzed the segmentation results from GPT-4o-mini and KNP and found that GPT-4o-mini did not always segment at the points identified by KNP. Since this was not a critical issue for aligning *bunsetsus* with the source segments, we decided to use GPT-4o-mini, which did not produce parsing errors on the dev set in our main experiments.

D Simul-COMET Scores

The Simul-COMET score results for `off-tgt` and `si-tgt` on the test set are shown in Table 11. Similar to the default COMET, i.e., `wmt22-comet-da`, Simul-COMET tended to assign higher scores to `off-tgt` than to `si-tgt` except for the mix-mono models. In addition, Simul-COMET tended to produce scores in a lower range compared to `wmt22-comet-da`.

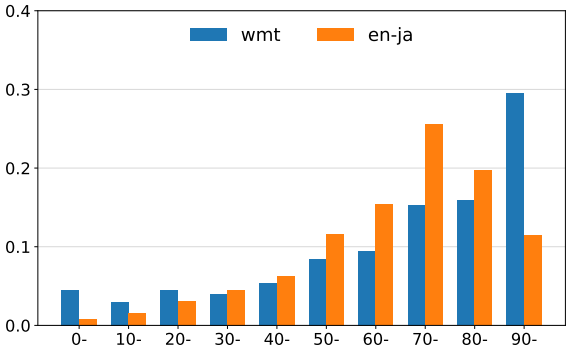

Figure 4: Quality score distributions of training data for `wmt22-comet-da` and Simul-COMET, denoted as wmt and en-ja, respectively. The number of instances is shown as proportions.

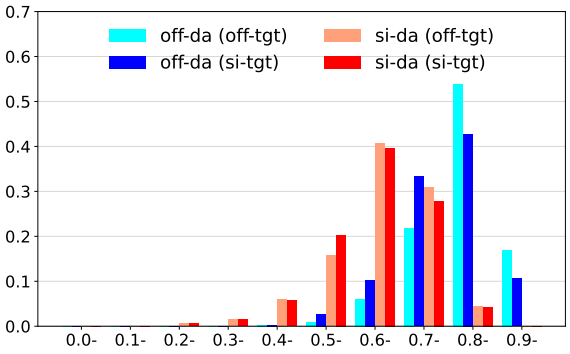

Figure 5: Simul-COMET score distributions in the off-da and si-da settings. off-tgt and si-tgt represent the type of target sentences. The number of instances is shown as proportions.

To examine these score range differences, we analyzed the score distributions of the training data for `wmt22-comet-da` and its en-ja subset, which served as the training data for Simul-COMET. Figure 4 shows that the en-ja data are distributed around a score of 70, whereas the COMET training data includes more samples in the higher-scoring range (mean=69.67, 68.68; SD=27.37, 20.09 for COMET and en-ja data, respectively).

After applying the monotonicity penalty, i.e., the training data for *-mono models, the score ranges naturally become lower: mean=63.68, 66.33; SD=20.40, 20.24 for off-mono and si-mono, respectively. The SDs are similar to that of the original en-ja data, indicating that the data distribution does not change substantially. These findings suggest that the differences in output score ranges arise from the score distributions used during training.

However, in the scratch setting, the `off-da` model produced scores in a higher range than

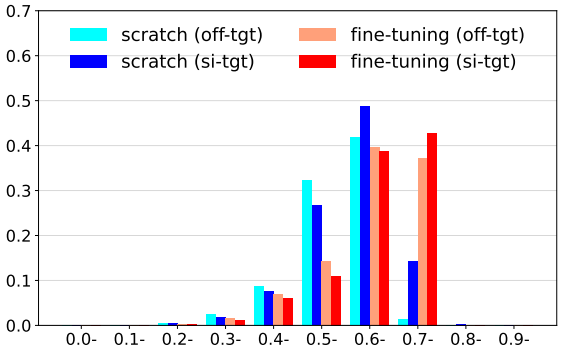

Figure 6: Simul-COMET score distributions in the mix-mono settings. scratch and fine-tuning denote models trained from scratch and fine-tuned from the default COMET model, respectively. off-tgt and si-tgt represent the type of target sentences. The number of instances is shown as proportions.

the `si-da` model, despite having identical quality scores in the training data, as shown in Figure 5. We examined whether differences in the target sentences themselves contribute to this by analyzing their PLL computed with XLM-R. We observed differences in PLL between `off-tgt` and `si-tgt`: PLL=−84.88, −91.83; length-normalized PLL=−2.60, −2.83, respectively, suggesting that XLM-R produces different sentence embeddings for `off-tgt` and `si-tgt`, which may influence the output score ranges of the trained Simul-COMET models.

In contrast, the `mix-mono` models assign higher scores to `si-tgt` than `off-tgt`, as shown in Figure 6. These trends further support the effectiveness of our strategies, which mix `off-tgt` and `si-tgt` and penalize quality scores based on monotonicity.

From these findings, it appears that differences in the target sentences themselves, as well as differences in the quality-score distributions of the training data, contribute to the observed differences in the Simul-COMET scores.

E MQM-based Guidelines

The rubrics used for human evaluation by professional simultaneous interpreters are shown in Table 13. In addition to these rubrics, we provided the interpreters with a set of guidelines that explained the purpose and background of the evaluation, the detailed evaluation procedures, and evaluation examples based on actual interpreted and translated utterances. Note that all the document given to the interpreters were in Japanese, but we present the English version of the rubrics in this study, which

Model	Rater A	Rater B	Avg. score
BLEU	0.0328	0.0613*	0.0389
wmt22-comet-da	0.0967	0.0701	0.1008*
mix-mono (scratch)	0.0647	0.0733	0.0687
mix-mono (fine-tuning)	**0.1099***	**0.1250***	**0.1295***

Table 12: Kendall's rank correlation coefficient between automatic metric scores and human evaluation scores. Avg. score indicates the correlation between automatic metric scores and the average of the scores provided by Rater A and Rater B. * indicates that the correlation coefficient was statistically significant ($p < 0.05$).

was translated with the assistance of DeepL.

F Correlation between Human Evaluation and Automatic Metric Scores

In addition to Spearman's rank correlation coefficient (Section 4.3), we also computed Kendall's rank correlation coefficient between automatic metric and human evaluation scores, as shown in Table 12. Similar to Spearman's rank correlation coefficient, BLEU exhibits substantially weaker correlation with human judgments than the COMET-based models. The `mix-mono (fine-tuning)` model achieved the highest correlation with the averaged human scores ($\tau = 0.1295$, $p < 0.05$). The `mix-mono` model also showed statistically significant correlations with the scores from both Rater A and B ($\tau = 0.1099, 0.1250$), $p < 0.05$, respectively). In contrast, the default COMET model, `wmt22-comet-da`, demonstrated a significant correlation with the averaged human score ($\tau = 0.1008$, $p < 0.05$), but did not with the individual rater scores ($\tau = 0.0967, 0.0701$, $p = 0.0531, 0.1701$ for Rater A and B, respectively). The correlations between `mix-mono (scratch)` model and human scores were not statistically significant ($\tau = 0.0647, 0.0733, 0.0687$, $p = 0.1959, 0.1515, 0.1545$, for Rater A, Rater B, and averaged score, respectively).

We further calculated correlation coefficients with different error weights and ranked the three models. Consistent with Spearman's rank correlation coefficient, the `mix-mono (fine-tuning)` model consistently ranked first, followed by the `wmt22-comet-da` model and the `mix-mono (scratch)` model.

Pearson's correlation coefficient is also commonly used to assess the relationship between automatic evaluation metrics and human evaluation scores. However, in this study, we adjusted the original DA scores by incorporating monotonicity, and the adjusted scores may exhibit different characteristics from the original ones. Since our focus is on the relative ranking of better or worse translations rather than precise numerical differences in evaluation scores, we used Spearman's rho and Kendall's tau instead of Pearson's r.

Severity	Criteria
Critical	The original meaning is significantly lost, and there is a possibility of serious misunderstanding. There is a possibility of serious risks (e.g., health damage, economic loss, damage to social reputation) in situations where decisions are made based on the translation.
Major	Errors where the original text's intent, logical flow, or meaning is miscommunicated or may be misunderstood due to omissions, additions, or mistranslations. Even considering the constraints of simultaneous interpretation, there is a risk of conveying incorrect understanding to the listener.
Minor	The main points and intent of the original text are conveyed, but there are inappropriate omissions, additions, rephrasing, or shifts in meaning in parts of the translation. While acceptable as interpretation, there is room for improvement in situations where accuracy is prioritized.
None	Errors that do not significantly impact the understanding of the main idea of the original text. Omissions or simplifications that are naturally made in simultaneous interpreting and do not hinder overall understanding. A translation that is not classified as an error.

(a) Accuracy

Severity	Criteria
Critical	A translation containing grammatical or lexical errors that cause a breakdown in meaning, making the content incomprehensible or likely to be misunderstood.
Major	A translation that contains grammatical or lexical errors, requiring effort to understand the meaning or posing a risk of misunderstanding. Translation that is difficult to use as is
Minor	Does not affect understanding of meaning, but grammatical or lexical awkwardness may cause the listener to feel uncomfortable. Generally acceptable for interpretation.
None	Even if there are unnatural parts, they are sufficiently acceptable as interpretation and do not hinder the transmission of meaning or the listener's understanding. SI-specific omissions, rephrasing, and incomplete sentences are acceptable.

(b) Fluency

Severity	Criteria
Critical	The word order in the translation does not correspond to the original text, and it is a typical reverse translation. It is closer to a written offline translation than simultaneous interpretation.
Major	There are multiple chunks in the translation that differ from the order in which they appear in the original text, or there are chunks that differ significantly in order. The timing of the translation is awkward as simultaneous interpretation (e.g., too much delay).
Minor	It is generally a monotonic translation, but contains minor reverse translation (non-monotonic word order) in some parts. The positions of multiple chunks are slightly different from the original text, but the overall flow is monotonic.
None	The order of information presentation in the translation is generally consistent with the chunk order of the original text, making it a natural monotonic translation. Even if there are minor rearrangements, they are limited to one instance between adjacent chunks.

(c) Monotonicity

Table 13: Rubrics for human evaluation