Tie-Calibrated COMETKiwi for Speech Translation Quality Estimation: IWSLT 2026 Metrics Track

Mubashir Hussain Shah **Aymen Fatima** **Kiho Choi*** **Daehee Jang***
Kyung Hee University, Republic of Korea
{shahmubashirhussain, fatimaaymen, aikiho, daehee87}@khu.ac.kr
*Corresponding authors

Abstract

We describe our submission to the IWSLT 2026 Speech Translation Metrics shared task, which targets reference-free quality estimation for English-to-German and English-to-Chinese speech translation. Our primary system combines COMETKiwi-22, applied to ASR transcripts, with a lightweight post-processing step called tie calibration: a learned score-bucketing that collapses near-identical scores into exact ties, reducing noisy within-document pairwise ranking errors. On the official development set the method achieves a segment-level Kendall τ_b of 39.4% on average, compared to 34.6% for plain COMETKiwi, 29.2% for SpeechQE, and 24.4% for BLASER 2.0 QE. System-level Soft Pairwise Accuracy is 88.0%, comparable to COMETKiwi (89.4%) and above SpeechQE (86.0%). The method requires no audio, no retraining, and one hyperparameter per target language tuned entirely on the training split.

1 Introduction

Speech translation (ST) evaluation is a challenging open problem. Unlike text machine translation (MT), ST outputs arise from automatically segmented audio streams, introducing segmentation mismatches and noise that degrade reference-based metrics (Sperber et al., 2024). Reference-free quality estimation (QE), which predicts translation quality without any reference, offers a practical alternative (Zerva et al., 2022). Yet dedicated QE research for speech translation is extremely sparse: virtually all QE work targets text MT, and applying text QE to ST requires treating ASR transcripts as clean source text, silently discarding the uncertainty introduced by the speech-to-text step.

The IWSLT 2026 Speech Translation Metrics shared task (Adelani et al., 2026) is the first dedicated evaluation campaign for ST QE,[1] evaluating systems on segment-level Kendall τ_b and system-level Soft Pairwise Accuracy (SPA) (Deutsch et al., 2023; Thompson et al., 2024).

Symbol	Meaning
s	Raw COMETKiwi score for a segment
s'	Calibrated score after bucketing
ϵ	Bucketing width (hyperparameter)
σ	Std. dev. of train-split scores
k	Scaling factor; $\epsilon = k \cdot \sigma$
$\mathcal{G}$	Grid of ϵ candidates
$\mathbf{s}^{\text{tr}}$	COMETKiwi scores on the training split
$\mathbf{s}^{\text{te}}$	Scores to calibrate (dev or test split)
$\mathbf{y}^{\text{tr}}$	Human DA labels on the training split
τ_b	Kendall τ_b (segment-level)
SPA	Soft Pairwise Accuracy (system-level)

Table 1: Notation used in this paper.

Our approach starts from a simple observation: COMETKiwi (Rei et al., 2022), applied to ASR transcripts, already produces a strong quality signal, but its continuous scores create many near-identical comparisons within documents. Since τ_b penalises every misordered pair equally regardless of magnitude, these arbitrary orderings inflate the apparent error rate without reflecting true quality differences. We address this with *tie calibration*: a learned ϵ-bucketing that rounds nearby scores to the same value, converting noisy near-ties into exact ties that τ_b ignores, with ϵ tuned on the training split alone and requiring no audio access, no fine-tuning, and no additional model.

Our contributions are: (i) a training-free post-processing step that improves τ_b by 4.8 points over COMETKiwi with no audio or retraining; (ii) ablations showing ϵ-bucketing improves segment ranking without degrading system-level SPA; and (iii) an analysis tracing the larger en-zh gain to tighter within-document score distributions.

[1] https://iwslt.org/2026/metrics

Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026), pages 318–322

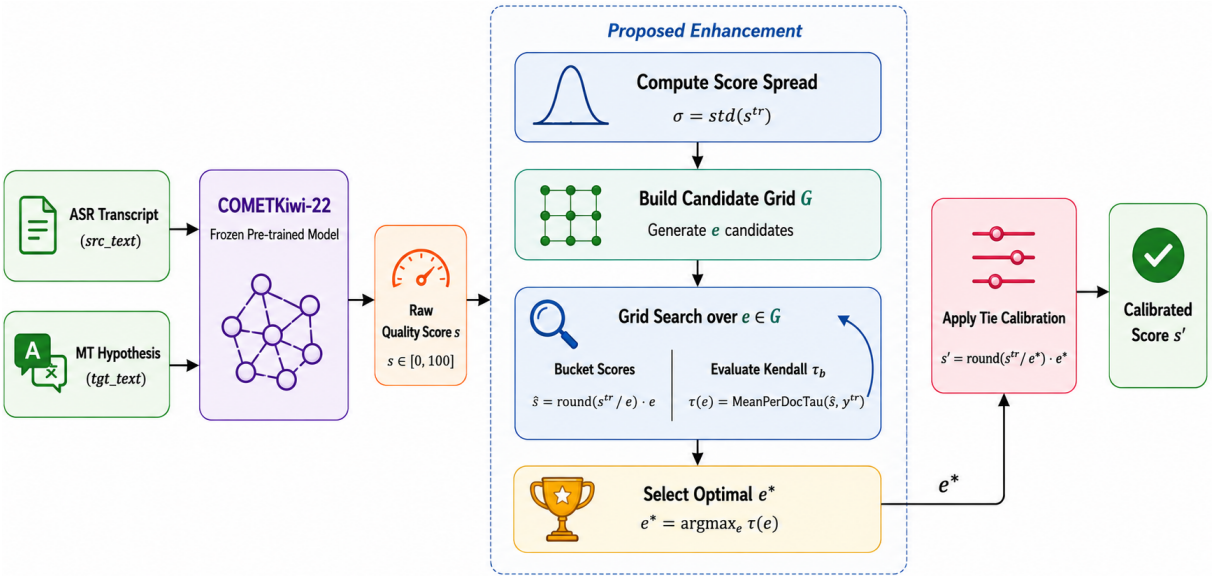

Figure 1: Overview of the tie-calibrated QE pipeline.

2 Task Description

The IWSLT 2026 Speech Translation Metrics task asks participants to predict a single real-valued quality score for each (source audio, ASR transcript, MT hypothesis) triple, without access to any reference translation. Two evaluation scenarios are considered: one may use the audio directly or rely solely on the ASR transcript.

The official evaluation metrics are segment-level Kendall τ_b, which for each source segment ranks all MT hypotheses and measures concordance with the human ranking averaged over all source segments, and system-level SPA, which measures how often the predicted system ranking agrees with the human ranking, weighting each pair by the statistical significance of the human preference (Thompson et al., 2024).

The shared task releases human-annotated data from IWSLT 2023, WMT 2024, and WMT 2025 for training, and IWSLT 2025 ACL Talks for development (Adelani et al., 2026).[2] The test set contains 48,044 EN→{DE, ZH} examples across five domains: ACL talks (64.5%), TV series (12.5%), call centre (11.3%), YouTube (8.1%), and business news (3.7%) (Adelani et al., 2026).[3] Each segment includes source speech and a `whisper-large-v3` transcript; ACL segments additionally provide human source transcriptions.

[2] https://huggingface.co/datasets/maikezu/iwslt2026-metrics-shared-train-dev

[3] https://speechm.cloud.cyfronet.pl/

Algorithm 1 Tie-calibration post-processing

Require: Training QE scores $\mathbf{s}^{\mathrm{tr}}$, human labels $\mathbf{y}^{\mathrm{tr}}$, scores to calibrate $\mathbf{s}^{\mathrm{te}}$
Ensure: Calibrated scores $\mathbf{s}'^{\mathrm{te}}$

// Step 1: Estimate score spread
$\sigma \leftarrow \mathrm{std}(\mathbf{s}^{\mathrm{tr}})$
Build $\mathcal{G}$ via Equation (2)
// Step 2: Grid search on training split
for each $\epsilon \in \mathcal{G}$ **do**
  $\tilde{\mathbf{s}} \leftarrow \mathrm{round}(\mathbf{s}^{\mathrm{tr}}/\epsilon) \cdot \epsilon$ *bucket scores*
  $\tau(\epsilon) \leftarrow \mathrm{MeanPerDocTau}(\tilde{\mathbf{s}}, \mathbf{y}^{\mathrm{tr}})$ *train only*
end for
// Step 3: Select and apply
$\epsilon^* \leftarrow \arg\max_{\epsilon \in \mathcal{G}} \tau(\epsilon)$
$\mathbf{s}'^{\mathrm{te}} \leftarrow \mathrm{round}(\mathbf{s}^{\mathrm{te}}/\epsilon^*) \cdot \epsilon^*$
return $\mathbf{s}'^{\mathrm{te}}$

3 System

Figure 1 illustrates the pipeline. Table 1 summarises all symbols used in the following sections.

3.1 Base Metric: COMETKiwi-22

COMETKiwi-22 (Rei et al., 2022)[4] is a reference-free QE model that scores a (source, hypothesis) pair using a predictor-estimator architecture built on a multilingual encoder, producing a scalar quality estimate in $[0, 1]$.

We use it as our base scorer in a text-only setting: the ASR transcript (`src_text`) serves as `src` and the system hypothesis (`tgt_text`) as `mt`, with no audio accessed. Weights are frozen; scores are rescaled to $[0, 100]$ before tie calibration.

[4] https://huggingface.co/Unbabel/wmt22-cometkiwi-da

System	Segment τ_b (%)				System SPA (%)			
	en-de	en-zh	Avg	$\Delta\tau_b$	en-de	en-zh	Avg	ΔSPA
Task-provided baselines								
COMETKiwi (Rei et al., 2022)	32.6	36.5	34.6	—	86.2	92.6	**89.4**	—
COMET-partial (Zouhar et al., 2026)	11.3	12.0	11.6	−23.0	44.4	68.7	56.6	−32.8
BLASER 2.0 QE (Seamless Communication, 2023)	22.0	26.8	24.4	−10.2	86.0	67.7	76.9	−12.5
SpeechQE (Han et al., 2024)	26.6	31.8	29.2	−5.4	78.6	**93.4**	86.0	−3.4
Ours								
C1: COMETKiwi, reproduced	32.9	36.6	34.7	+0.1	86.3	89.1	87.7	−1.7
Primary: per-lang ϵ^*, round	**35.6**	**43.2**	**39.4**	**+4.8**	**86.3**	89.8	88.0	**−1.4**

Table 2: Development-set results (official evaluation scripts). Δ values are relative to the COMETKiwi baseline. Task-provided baseline values are taken directly from the organiser repository; minor deviations from our reproduced C1 reflect environment differences. **Bold**: best value in each column.

3.2 Tie Calibration

Background. Kendall's τ_b is a rank-correlation coefficient computed by comparing all pairs of items within a group and counting concordant minus discordant pairs, normalised by the total number of comparable pairs (Deutsch et al., 2023). Tied pairs contribute zero to both numerator and denominator: a tie is neither concordant nor discordant. A scorer assigning slightly different continuous scores to effectively indistinguishable hypotheses produces many arbitrary micro-orderings; each misordering counts as a discordant pair and depresses τ_b even though it reflects no true quality distinction. Deutsch et al. (2023) formalise this weakness and propose a family of tie-calibration procedures to mitigate it.

Method. We apply the following transformation to each raw score s:

$$s' = \text{round}\left(\frac{s}{\epsilon}\right) \cdot \epsilon \quad (1)$$

where $\epsilon > 0$ is the bucketing width. Any two scores within $\epsilon/2$ collapse to the same bucket value s'. We also evaluate `floor` and `ceil` variants, denoted $\lfloor s/\epsilon \rfloor \cdot \epsilon$ and $\lceil s/\epsilon \rceil \cdot \epsilon$, respectively (Section 4.2).

Selecting ϵ. Let σ denote the standard deviation of COMETKiwi scores on the training split for a given language pair. We search over:

$$\mathcal{G} = \left\{ k{\cdot}\sigma \;:\; k \in \{0,\ 0.0025,\ 0.005,\ \ldots,\ 0.20\} \right\} \quad (2)$$

For each $\epsilon \in \mathcal{G}$ we bucket the training-split scores, compute mean per-document τ_b against human annotations, and retain the ϵ^* that maximises this value. The development set is never used during selection. We restrict the search to DE-target and ZH-target rows of the training pool, tuning ϵ separately per target language. Algorithm 1 states the full procedure. The selected values are $\epsilon^* = 3.497$ (en-de) and $\epsilon^* = 2.756$ (en-zh), both at $k = 0.20$. The difference reflects a language-specific property: COMETKiwi assigns tighter within-document score distributions to Chinese hypotheses, so a smaller ϵ suffices.

3.3 Contrastive Submissions

Three contrastive systems accompany the primary: C1 (plain COMETKiwi, no calibration), C2 (shared ϵ across both language pairs, round bucketing), and C3 (per-language ϵ, floor bucketing).

4 Results

4.1 Comparison with Baselines

Table 2 reports development-set results for our primary system against the baselines, using the provided evaluation scripts.[5] Segment-level τ_b is computed per document and averaged; SPA uses the statistical-significance weighting of Thompson et al. (2024).[6]

The primary system achieves 39.4% mean τ_b, a gain of +4.8 points over the task-provided COMETKiwi (34.6%) and +10.2 over SpeechQE (29.2%). The improvement is consistent across both language pairs: +3.0 for en-de and +6.7 for en-zh. System-level SPA is 88.0%, which is 1.4 points below the task-provided COMETKiwi (89.4%) but 2.0 above SpeechQE (86.0%). The modest SPA gap is expected and discussed in Section 5.

[5] https://github.com/zouharvi/iwslt26-metrics

[6] All inference runs on a single NVIDIA RTX A6000 GPU with batch size 16.

Bucketing variant	τ_b avg (%)	SPA avg (%)
C1: none	34.7	87.7
C2: shared ϵ, round	38.5	**88.8**
C3: per-lang, floor	37.8	87.9
per-lang, ceil	37.8	87.9
Primary: per-lang, round	**39.4**	88.0

Table 3: Ablation of bucketing function and ϵ-sharing strategy (development set). Scores averaged over both language pairs.

k	τ_b (%)		SPA (%)	
	en-de	en-zh	en-de	en-zh
0.00	32.9	36.6	86.3	89.1
0.01	33.1	37.3	86.3	89.2
0.03	33.3	38.5	86.3	89.7
0.05	33.4	38.9	86.3	88.2
0.10	34.4	41.0	86.3	**90.9**
0.15	34.7	42.9	86.2	87.9
0.20	**35.6**	**43.2**	**86.3**	89.8

Table 4: τ_b and SPA across scaling factors k (per-language ϵ, round bucketing). k=0 is plain COMETKiwi (C1).

4.2 Ablation Studies

Bucketing strategy. Table 3 isolates the effect of the bucketing function, keeping the per-language ϵ^* fixed. Round bucketing yields the highest τ_b. Shared ϵ (C2) improves SPA by 0.8 points at the cost of lower τ_b. Floor and ceil are symmetric and intermediate.

Epsilon sensitivity. Table 4 reports τ_b and SPA across the full range of scaling factors k (where $\epsilon = k \cdot \sigma$) under per-language round bucketing. τ_b increases monotonically from $k = 0$ to $k = 0.20$; for en-zh the gradient is shallower above $k = 0.15$, while for en-de the increments remain small throughout, placing the optimal $k = 0.20$ in a stable region for both language pairs. SPA remains within two percentage points of C1 throughout the sweep, confirming that tie calibration is safe to deploy without risking system-level performance.

5 Analysis

Why tie calibration improves τ_b but not SPA. τ_b is computed within documents where hypotheses from 8 or more MT systems may be of near-equal quality; COMETKiwi assigns them near-equal but non-identical scores, producing arbitrary orderings that count as discordant pairs. Tie calibration collapses these into exact ties, which τ_b ignores. At the system level, inter-system score differences are large enough that bucketing does not change system rankings, leaving SPA unchanged — as theoretically expected (Deutsch et al., 2023).

Why en-zh benefits more. COMETKiwi produces tighter within-document score distributions for Chinese than for German, as shown in Table 5. A tighter distribution means more hypothesis pairs fall within the bucketing radius $\epsilon/2$, so tie calibration converts more near-ties into exact ties and yields a larger τ_b gain. This reflects a property of COMETKiwi's scoring behaviour on these language pairs and may not generalise to other QE models.

Metric	en-de	en-zh
Mean within-doc σ (COMETKiwi)	8.31	5.49
Median within-doc σ (COMETKiwi)	7.63	3.15
τ_b gain from calibration (%)	+2.7	+6.7

Table 5: Within-document COMETKiwi score spread (development set) and corresponding τ_b gains from tie calibration.

Why our SPA is slightly below the task-provided COMETKiwi. The task-provided COMETKiwi baseline achieves 89.4% SPA while our reproduced C1 achieves 87.7%. The 1.7-point gap likely reflects environment differences (package versions, floating-point precision), as confirmed by nearly identical τ_b values (34.6% vs. 34.7%). Tie calibration then raises our SPA to 88.0%, narrowing the gap to 1.4 points.

Text-only versus audio-based approaches. Although Han et al. (2024) showed that E2E audio QE can outperform cascaded text QE by avoiding ASR information loss, the SpeechQE baseline (29.2%) is 10.2 points below our text-only system on the development set. The development data consists of ACL Talks with high-quality, scripted academic speech and human-corrected transcripts, which may advantage text-only methods. The broader test-set evaluation (Adelani et al., 2026) confirms this picture: contrary to the premise of the task, text-based metrics match or outperform their multimodal counterparts overall, suggesting that transcription quality is not the primary bottleneck for QE performance on this benchmark.

Applicability to other QE models. Tie calibration is model-agnostic: any scorer producing continuous scores can benefit, provided ϵ is tuned on representative annotated data and within-document score variance is sufficient for bucketing to be meaningful.

6 Conclusion

Tie calibration, a learned ϵ-bucketing applied as post-processing to frozen COMETKiwi scores, improves segment-level τ_b by 4.8 points over the task-provided COMETKiwi baseline and 10.2 points over SpeechQE on the IWSLT 2026 development set, while preserving system-level SPA. The gain is larger for en-zh (+6.7) than en-de (+3.0), explained by tighter within-document score distributions for Chinese (mean $\sigma = 5.49$ vs. 8.31 for German). The method requires no audio, no retraining, and one hyperparameter per target language tuned entirely on the training split.

Limitations

Our text-only design means transcription errors propagate silently: if the ASR output misrepresents the source, COMETKiwi may assign a high score to an incorrect translation with no way to detect the failure. The bucketing width ϵ is also corpus-specific and must be re-tuned when applied to a new language, domain, or QE model.

Acknowledgements

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2025-16069081).

References

David Ifeoluwa Adelani, Victor Agostinelli, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Sebastien Bratières, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Lizhong Chen, Marcello Federico, Marco Gaido, Mahendra Gupta, HyoJung Han, Ali Hatami, David Javorský, Yejin Jeon, Marek Kasztelnik, Antoine Laurent, and 33 others. 2026. Speech translation and metrics in 2026: Findings of the iwslt campaign. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, California, US. Association for Computational Linguistics.

Daniel Deutsch, George Foster, and Markus Freitag. 2023. Ties matter: Meta-evaluating modern metrics with pairwise accuracy and tie calibration. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12914–12929, Singapore. Association for Computational Linguistics.

HyoJung Han, Kevin Duh, and Marine Carpuat. 2024. SpeechQE: Estimating the quality of direct speech translation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21852–21867, Miami, Florida, USA. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Seamless Communication. 2023. SeamlessM4T: Massively multilingual & multimodal machine translation. *Preprint*, arXiv:2308.11596.

Matthias Sperber, Ondřej Bojar, Barry Haddow, Dávid Javorský, Xutai Ma, Matteo Negri, Jan Niehues, Peter Polák, Elizabeth Salesky, Katsuhito Sudoh, and Marco Turchi. 2024. Evaluating the IWSLT2023 speech translation tasks: Human annotations, automatic metrics, and segmentation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6484–6495, Torino, Italia. ELRA and ICCL.

Brian Thompson, Nitika Mathur, Daniel Deutsch, and Huda Khayrallah. 2024. Improving statistical significance in human evaluation of automatic metrics via soft pairwise accuracy. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1222–1234, Miami, Florida, USA. Association for Computational Linguistics.

Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. 2022. Findings of the WMT 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Vilém Zouhar, Maike Züfle, Beni Egressy, Julius Cheng, Mrinmaya Sachan, and Jan Niehues. 2026. Early-exit and instant confidence translation quality estimation. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 55–76, Rabat, Morocco. Association for Computational Linguistics.