Pairwise Ranking Fine-tuning of CometKiwi for Speech Translation Quality Estimation

Pranav Gupta
Cisco

Abstract

We describe our submission to the IWSLT 2026 Speech Translation Metrics Shared Task. The task requires reference-free quality estimation of speech translation outputs, evaluated by segment-level Kendall's τ computed within documents. We fine-tune CometKiwi-22, a 580M-parameter quality estimation model, with a pairwise ranking objective that directly optimizes the evaluation criterion. By constructing within-document translation pairs and training with an adaptive margin ranking loss combined with MSE calibration, our system achieves 35.2% per-source Kendall's τ on the development set, improving over the organizers' CometKiwi-22 baseline of 34.6%. We also report results from several other metrics including MetricX-24, xCOMET-XL, BLASER-2, and a LightGBM ensemble, finding that the single pairwise-trained model outperforms ensemble approaches due to train-test signal circularity constraints.

1 Introduction

Quality estimation (QE) for machine translation aims to predict translation quality without access to reference translations (Specia et al., 2018). While QE has been extensively studied for text-to-text translation, evaluating speech translation systems poses additional challenges: automatically segmented speech lacks clean sentence boundaries, ASR transcripts contain recognition errors, and the distribution of translation quality varies across documents and systems.

The IWSLT 2026 Metrics Shared Task frames this as a reference-free quality estimation problem: given an ASR transcript of the source speech and a system translation, predict a quality score that correlates with human judgments. Submissions are evaluated using per-document Kendall's τ_b averaged across documents, which measures the ability of a metric to correctly rank translations within the same source document.

This evaluation setup creates a mismatch with standard QE training objectives. Models like CometKiwi (Rei et al., 2022) are trained with mean squared error (MSE) loss to predict absolute quality scores, but the evaluation rewards correct *relative orderings* within documents rather than accurate absolute predictions. We exploit this mismatch by fine-tuning with a pairwise ranking loss that directly optimizes within-document ordering.

2 Related Work

The dominant paradigm for training learned MT metrics uses MSE regression on direct assessment (DA) or MQM scores (Rei et al., 2022). While pairwise and listwise ranking losses are well-established in learning-to-rank (Burges, 2010), their application to MT metric training remains limited. Concurrently, Züfle et al. (2025) propose COMET-poly, which provides alternative translations as input context at inference time to enable comparative scoring—a complementary approach that modifies the model's *input* rather than its *training objective*. Our work instead modifies the loss function to directly optimize ranking, which requires no additional candidates at inference time and is applicable to any regression-based metric.

3 Task Description

The IWSLT 2026 Metrics Shared Task (Adelani et al., 2026) provides training data comprising human quality annotations from IWSLT 2023, WMT 2024, and WMT 2025 (33,721 segments across 17 language pairs), and a development set from IWSLT 2025 ACL Talks (5,556 segments, English–German and English–Chinese). The test set contains 48,044 segments (24,016 *en–de*, 24,028 *en–zh*) from IWSLT 2026.

Each segment includes an ASR transcript (`src_text`), a system translation (`tgt_text`), a document identifier (`doc_id`), and a human qual-

Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026), pages 332–335

ity score. Segments are grouped by document, and evaluation computes Kendall's τ_b within each document, then averages across documents and language pairs.

4 System Description

4.1 Base Model

Our primary submission is based on CometKiwi-22 (Rei et al., 2022), a quality estimation model built on XLM-RoBERTa-Large (Conneau et al., 2020) with approximately 580M parameters. CometKiwi-22 was trained on direct assessment data from WMT shared tasks and uses source and machine translation as input (no reference required), making it suitable for the QE setting. The pretrained model achieves 34.6% per-source Kendall's τ on the development set as reported by the organizers.

4.2 Pairwise Ranking Loss

Standard QE fine-tuning minimizes MSE between predicted and gold scores. However, Kendall's τ measures the fraction of concordant pairs—it is invariant to monotonic transformations of the predicted scores. A model with poor absolute predictions but correct pairwise orderings achieves perfect τ.

We construct training pairs (x_i^+, x_i^-) from translations of the same source document where x_i^+ has a higher human score than x_i^- by at least 1 point (on a 0–100 scale). From the 11,279 English–German and English–Chinese training segments, this yields 103,062 pairs, which we subsample to 50,000.

Our loss combines an adaptive margin ranking loss with MSE:

$$\mathcal{L} = \lambda \cdot \mathcal{L}_{\text{MSE}} + (1 - \lambda) \cdot \mathcal{L}_{\text{rank}} \tag{1}$$

where $\lambda = 0.3$ and:

$$\mathcal{L}_{\text{rank}} = \max\left(0,\ m_i - (s_i^+ - s_i^-)\right) \tag{2}$$

The adaptive margin $m_i = \text{clamp}(\Delta_i/2, \text{min} = 0.01)$ scales with the gold score difference Δ_i, so pairs with larger quality gaps demand proportionally larger predicted separations. This prevents the model from satisfying the loss with uniformly small score differences, encouraging well-calibrated rankings even among translations of similar quality.

The MSE component preserves score interpretability and prevents representation collapse, acting as a regularizer.

4.3 Training Details

The original CometKiwi-22 was trained end-to-end from the start using MSE plus word-level cross-entropy on DA annotations for 2 epochs (Rei et al., 2022). Our fine-tuning departs in two ways: (1) we replace the loss with a pairwise ranking objective, and (2) we use a two-phase schedule that initially freezes the encoder:

- **Phase 1** (first 30% of epoch 1): The XLM-RoBERTa encoder is frozen; only the regression head is trained (learning rate 1×10^{-5}). This allows the head to adapt to the new pairwise loss before the encoder representations shift.
- **Phase 2** (remainder): The encoder is unfrozen with a lower learning rate (5×10^{-7}) while the head continues at 1×10^{-5}.

The 30% freeze duration was chosen to cover approximately one full pass through the paired training data before encoder adaptation begins; a similar gradual unfreezing strategy is standard in transfer learning (Howard and Ruder, 2018). The loss weight $\lambda = 0.3$ was selected to prioritize ranking (70% of the gradient signal) while retaining enough MSE regularization to prevent score collapse. We did not perform a full grid search due to computational constraints; preliminary experiments with $\lambda \in \{0.1, 0.3, 0.5\}$ showed 0.3 performed best on the development set, though the differences were within 0.3 τ points.

We train for up to 10 epochs with batch size 32, gradient clipping at 1.0, cosine learning rate decay with 10% warmup, and early stopping with patience 3 based on development set per-source τ. Training completes in approximately 30 minutes on a single GPU.

4.4 Contrastive Systems

In addition to our primary pairwise-trained submission, we explored several contrastive approaches during development:

MSE Fine-tuning. Standard fine-tuning of CometKiwi-22 with MSE loss on the training set gold scores. This achieved 35.1% per-source τ, comparable to pairwise training but without the direct ranking optimization.

Pretrained Metrics. We evaluated several pretrained metrics without fine-tuning: CometKiwi-23-XXL (10.7B parameters, 29.8% τ), xCOMET-

Method	en–de	en–zh	Avg
Organizers' baselines			
COMET-partial	11.3	12.0	11.6
BLASER-2 QE	22.0	26.8	24.4
SpeechQE	26.6	31.8	29.2
CometKiwi-22	32.6	36.5	34.6
Our pretrained evaluations			
BLASER-2 QE (text)	28.0	31.8	27.1
xCOMET-XL	30.5	29.9	29.2
CometKiwi-23-XXL	32.1	35.3	29.8
MetricX-24-Hybrid	29.6	31.7	29.9
CometKiwi-22	33.0	36.6	32.3
Our fine-tuned / combined			
CK-22 MSE fine-tuned	33.5	39.8	35.1
CK-22 pairwise (primary)	34.1	39.4	35.2
LightGBM ensemble	34.8	39.4	35.9

Table 1: Segment-level per-source Kendall's τ (%) on the IWSLT 2026 development set. Our primary submission is CometKiwi-22 fine-tuned with pairwise ranking loss.

XL (29.2% τ), MetricX-24-Hybrid-XXL (29.9% τ), and BLASER-2 QE (27.1% τ). Notably, the much larger CometKiwi-23-XXL performed *worse* than CometKiwi-22 (32.3% τ) on this task, suggesting that model scale alone does not improve within-document ranking for speech translation QE.

LightGBM Ensemble. We trained a LightGBM model on the scored training set to combine signals from multiple pretrained metrics. The ensemble achieved 35.9% τ on development, the highest overall. However, the ensemble could only use pretrained metric signals—not the fine-tuned model scores, since those models were trained on the same data. This circularity constraint limited the ensemble to combining weaker individual signals, and its advantage over the single pairwise model was marginal.

5 Results

Table 1 presents per-source Kendall's τ on the development set for all methods we evaluated.

5.1 Analysis

Pairwise loss vs. MSE. Both fine-tuning approaches improve substantially over the pretrained baseline (+2.9 and +2.6 points respectively). The pairwise model shows a small overall advantage, though the MSE model is slightly better on en–zh. The two losses optimize different objectives: pairwise training directly improves ranking, while MSE incidentally improves ranking by reducing absolute prediction error.

Model scale is not sufficient. CometKiwi-23-XXL (10.7B parameters) achieved lower per-source τ than CometKiwi-22 (580M) despite being 18$\times$ larger. This suggests that the task—ranking speech translation quality within documents—benefits more from task-specific fine-tuning than from scale of the pretrained metric.

Ensemble circularity. The LightGBM ensemble achieved the highest development τ (35.9%), but could only be trained on pretrained metric signals. Fine-tuned model scores cannot serve as ensemble features when the fine-tuned model was trained on the same data, as this introduces data leakage. We submitted the single pairwise model as our primary system because it avoids this circularity and its performance is within 0.7 points of the ensemble.

6 Conclusion

We presented a simple and effective approach to speech translation quality estimation: fine-tuning CometKiwi-22 with pairwise ranking loss that directly optimizes the evaluation criterion. The method achieves 35.2% per-source Kendall's τ on the development set, a modest improvement over the 34.6% baseline. Our experiments suggest that for within-document ranking tasks, directly optimizing the ranking objective with a moderate-sized model is more effective than scaling to larger pretrained metrics.

Limitations

Our approach relies entirely on ASR transcripts and does not use the source audio signal, which could provide additional information about translation quality (e.g., prosody, disfluencies, speaker characteristics). We also did not explore document-level context—each segment is scored independently, though the evaluation groups segments by document. Incorporating document context could help the model make better relative judgments.

We fine-tune using only English–German and English–Chinese training data (the two develop-

ment/test language pairs), discarding the remaining 15 language pairs. This decision was motivated by the pairwise training setup: pairs are constructed *within documents*, and cross-lingual pairs would conflate language-specific quality distributions with true ranking signal. While COMET-family models benefit from multilingual pretraining, our ranking fine-tuning operates at a different level—optimizing relative orderings within a language pair—where out-of-distribution language pairs could introduce noise. Future work should investigate whether auxiliary language pairs improve generalization, particularly in a curriculum or multitask setup that separates cross-lingual transfer from within-document ranking.

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.

Christopher J. C. Burges. 2010. From RankNet to LambdaRank to LambdaMART: An overview. In *Microsoft Research Technical Report MSR-TR-2010-82*.

Alexis Conneau, Karttikeya 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. Association for Computational Linguistics.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 328–339. Association for Computational Linguistics.

Ricardo Rei, José de Souza Tomé, Duarte Fernandes, Chrysoula Zerva, Tharindu Ranasinghe, Craig de Souza, Pierre Colombo, 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. Association for Computational Linguistics.

Lucia Specia, Frédéric Blain, Varvara Logacheva, Ramón Astudillo, and André F. T. Martins. 2018. Findings of the WMT 2018 shared task on quality estimation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 689–709. Association for Computational Linguistics.

Maike Züfle, Vilém Zouhar, Tu Anh Dinh, Felipe Maia Polo, Jan Niehues, and Mrinmaya Sachan. 2025. COMET-poly: Machine translation metric grounded in other candidates. In *Proceedings of the Tenth Conference on Machine Translation (WMT)*. Association for Computational Linguistics.