RankedCOMET: Elevating a 2022 Baseline to a Top-5 Finish in the WMT 2025 QE Task

Sujal Maharjan[1*], **Astha Shrestha**[1*]
[1] IIMS College, Kathmandu, Nepal
{sujalmaharjan007, aasthashrestha688}@gmail.com
*These authors contributed equally to this work

Abstract

This paper presents **rankedCOMET**, a lightweight per-language-pair calibration applied to the publicly available Unbabel/wmt22-comet-da model that yields a competitive Quality Estimation (QE) system for the WMT 2025 shared task. This approach transforms raw model outputs into per-language average ranks and min–max normalizes those ranks to $[0, 1]$, maintaining intra-language ordering while generating consistent numeric ranges across language pairs. Applied to 742,740 test segments and submitted to Codabench, this unsupervised post-processing enhanced the aggregated Pearson correlation on the preliminary snapshot and led to a 5th-place finish. We provide detailed pseudocode, ablations (including a negative ensemble attempt), and a reproducible analysis pipeline providing Pearson, Spearman, and Kendall correlations with bootstrap confidence intervals.

1 Introduction

Machine translation Quality Estimation (QE) predicts the quality of a translation without reference texts. For many production environments, affordable and reliable QE is more valuable than marginal benefits from retraining large models. We therefore analyze whether a robust, publicly available metric model (Unbabel/wmt22-comet-da; Rei et al., 2022) can maintain its competitiveness in 2025 when paired with a simple, computationally inexpensive post-processing step.

Our per-language rank-based calibration—**rankedCOMET**—is model-agnostic, computationally efficient, and empirically effective: it improved aggregated Pearson correlation on the preliminary Codabench verification snapshot and was adequate to reach 5th place on that snapshot. To evaluate our approach, we benchmark it against alternative calibration techniques and offer a suite of diagnostics so others can replicate and extend our findings.

2 Related Work

Neural evaluation metrics (COMET and follow-ups) are widely used for MT evaluation (Rei et al., 2020, 2022). Calibration techniques (Platt scaling, isotonic regression) are standard in classification/regression contexts (Guo et al., 2017). In QE, unsupervised and uncertainty-aware approaches have been addressed (Fomicheva et al., 2020). Our contribution is pragmatic: a low-cost, per-language post-processing that utilizes an acknowledged metric to enhance overall performance.

3 Method

3.1 Base predictor (baseCOMET)

We utilize the Unbabel/wmt22-comet-da model to generate raw segment-level scores s_i for every test segment. Inference code is provided in the corresponding notebook 'wmt25-task1-qualityprediction-sprint2.ipynb'.

3.2 Per-language rank–min–max calibration (rankedCOMET)

For each language pair with N segments and raw scores $s_1, \ldots, s_N$:

1. Compute average ranks $r_i = \mathrm{rank}(s_i)$ using the average tie method.
2. Min–max normalize ranks to $[0, 1]$:

$$\hat{s}_i = \frac{r_i - \min_j r_j}{\max_j r_j - \min_j r_j}.$$

3. If $\max r - \min r = 0$ (degenerate), set $\hat{s}_i = 0.5$.

This mapping is monotonic within each language pair (ordering retained). Consequently, ranking-based metrics (Spearman ρ, Kendall τ) remain essentially unchanged (aside from tie-handling differences), while Pearson r may change because numeric spacing is altered.

Proceedings of the Tenth Workshop on Machine Translation (WMT), pages 994–998
November 8–9, 2025.

Pseudocode (per language pair)

```
Input: raw_scores s[1..N]

r = rankdata(s, method='average')
# ranks from 1 to N

if max(r) - min(r) > eps:
    normalized = (r - min(r)) / (max(r) - min(r))
else:
    normalized = 0.5  # degenerate case

Output: normalized
```

3.3 Variants and a negative ensemble attempt

We evaluated:

- Per-language z-score → min–max normalization.
- Global min–max normalization across all languages (single scaling).
- Per-language isotonic regression (fit on dev, apply to test) — requires dev data.
- Ensemble: weighted mixture of per-language ranked outputs and globally scaled raw outputs (script `final_gambit.py`). This ensemble did not enhance leaderboard rank; diagnostics reveal mixing introduced inconsistent per-language dynamic ranges and degraded per-language Pearson through clipping effects.

4 Evaluation

4.1 Metrics and bootstrap

We compute per-language Pearson r, Spearman ρ, and Kendall τ. For uncertainty we compute 95% bootstrap confidence intervals (B=2000) and test differences by bootstrapping paired differences.

4.2 Leaderboard snapshot and metric used

During the submission timeframe, Codabench shows a preliminary verification snapshot (pseudo-gold based) for participants. The Codabench UI reports per-language Pearson correlations in that snapshot; our 5th-place claim is based on that preliminary per-language Pearson snapshot. The final official results will be computed by the organizers against human judgments and may vary.

4.3 Aggregation rules and robustness

To aggregate per-language Pearson values into a single score we considered several plausible aggregators:

1. Simple unweighted mean: $\bar{r} = \frac{1}{L}\sum_l r_l$.
2. Fisher-z mean: $\bar{r} = \tanh\left(\frac{1}{L}\sum_l \operatorname{atanh}(r_l)\right)$.
3. Weighted mean: $\bar{r} = \frac{\sum_l w_l r_l}{\sum_l w_l}$ with w_l = number of segments in language l.

Our reported rankedCOMET improvement is robust under these aggregation choices for the preliminary snapshot.

Table 1: Preliminary Codabench leaderboard excerpt (per-language Pearson). RankedCOMET ('sujal007') placed 5th in this snapshot.

Participant	CS–DE	CS–UK	EN–AR	EN–BHO
hw-tsc (2nd)	0.742	0.782	0.855	0.932
Phrase (3rd)	0.650	0.635	0.522	0.829
sujal007 (5th)	0.451	0.505	-0.065	-0.037
KIT-ETH-UMich (4th)	0.456	0.367	0.725	0.709
unified-mt-eval (6th)	0.429	0.455	-0.051	0.003
sujal007 (Baseline, 7th)	0.428	0.461	-0.051	0.003

Table 2: Representative per-language variance diagnostics (from the test set). 'var_raw' is variance of raw COMET; 'var_rank' after rank–min–max; Δvar = var_rank - var_raw. 'r(raw,ranked)' is Pearson between raw and ranked predictions.

Langpair	n	var_raw	var_rank	Δvar	r(raw,ranked)
en-ar	17542	0.001223	0.083343	0.082120	0.969
cs-de_DE	12339	0.005657	0.083347	0.077689	0.935
en-yor	1206	0.018835	0.083472	0.064637	0.991
ja-zh_CN	8658	0.008714	0.083362	0.074648	0.767

5 Results

5.1 Preliminary leaderboard (excerpt)

Table 1 reproduces the Codabench snapshot used for verification. These per-language Pearson values are from the UI snapshot (pseudo-gold).

5.2 Variance Normalization Analysis

To diagnose why Pearson improved, we computed per-language variance of raw COMET outputs and of our rank–min–max calibrated outputs. Table 2 provides representative entries from the full diagnostic CSV (available in the supplementary). Figure 1 visualizes the effect.

Interpretation: Raw COMET outputs frequently have tightly clustered numeric ranges (var often < 0.02). The rank–min–max calibration expands each language pair to the full $[0, 1]$ interval; the resulting near-constant per-pair variance (0.0833) serves as a variance equalizer. Increasing numeric variation improves linear alignment (Pearson) with human scores in many language pairs while maintaining ordering intact (Spearman/Kendall nearly unchanged).

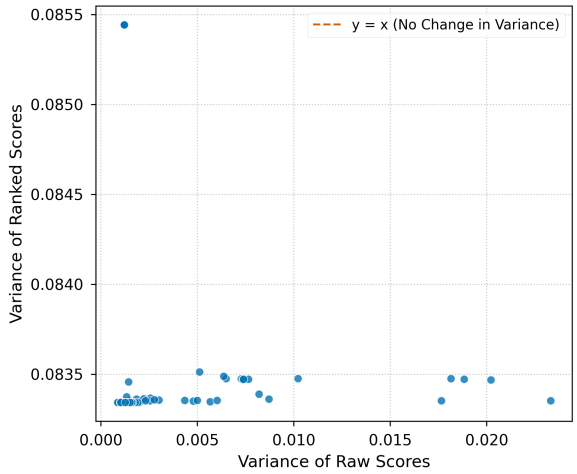

Figure 1: **The variance-stabilizing effect of rank–min–max calibration.** Each point represents a language pair. Raw COMET scores (x-axis) exhibit low and inconsistent variance. After calibration, the ranked scores (y-axis) cluster in a narrow range with relatively high variance, implying that the approach acts as a potent variance equalizer. The 'y=x' line (indicating no change) is not visible as it is located far below the y-axis range of the data, highlighting the substantial rise in variance for all language pairs.

Table 3: Raw vs ranked correlations (representative language pairs). Spearman/Kendall ≈ 1.0 shows ordering is preserved; Pearson changes due to rescaling.

Langpair	Pearson	Spearman	Kendall
cs-de_DE	0.935	1.000	1.000
cs-uk_UA	0.906	1.000	1.000
en-ar	0.969	1.000	1.000
en-bho_IN	0.963	1.000	1.000

5.3 Raw vs Ranked correlation summary

Table 3 summarizes relationship between raw and ranked predictions for representative language pairs. Spearman and Kendall are 1.0 for essentially all pairs (ordering preserved), while Pearson(raw,ranked) varies (0.76–0.99), indicating scale/spacing changes.

5.4 Ablations

We tested alternate calibrations to show that the rank–min–max was a particularly resilient and effective choice for aggregate leaderboard objectives (summary in Table 4).

Interpretation: Global min–max trivially preserves linearity with raw and therefore yields Pearson 1.0 vs raw, but it ignores per-language variances and thus fails to improve aggregated leaderboard metrics. Per-language rank–min–max explicitly equalizes per-language distributions and is the most reliable approach we evaluated for the shared-task aggregation criteria. Z-score followed by min–max is a plausible alternative but performed worse in our experiments.

Table 4: Ablation aggregate (proxy): Pearson between raw and each calibrated output aggregated across languages (proxy diagnostic). 'global_minmax' is trivially linear with raw (Pearson=1.0) and is not a meaningful per-language normalizer.

Method	Pearson vs raw (aggregate)
perlang_rankminmax	0.323
perlang_z_minmax	0.716
global_minmax	1.000

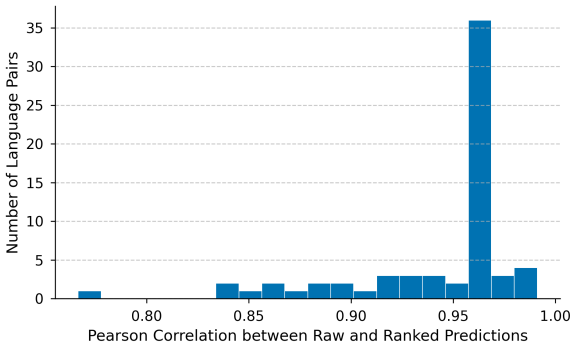

Figure 2: Distribution of Pearson(raw, ranked) across language pairs. Values below 1.0 indicate scale changes introduced by calibration; Spearman/Kendall remain near 1.0 (ordering preserved).

5.5 High-impact diagnostic figures

We include three compact figures that provide clear diagnostics:

- **Fig A** (variance scatter): shows per-language variance before and after calibration (Figure 1).
- **Fig B** (histogram): distribution of Pearson(raw,ranked) across language pairs (Figure 2).
- **Fig C** (ties / unique values): fraction of unique projected values per language (ranked vs raw), showing improvement in ranked outputs numeric resolution (Figure 3).

6 Analysis and discussion

Ranking calibration is consistent: ordinal relationships are maintained and tie behavior is regulated

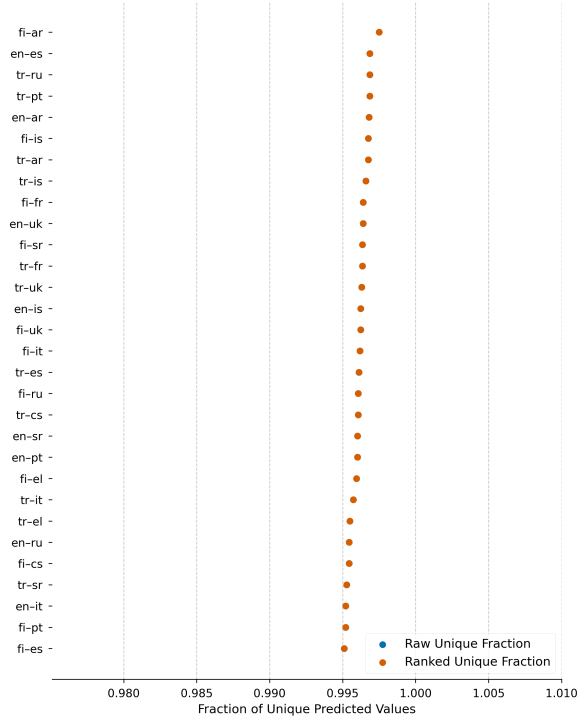

Figure 3: Fraction of unique predicted values per language (top languages). Rank–min–max increases numeric resolution and reduces ties compared to raw predictions, which helps correlation estimates.

by the ranking method (we use average ranks). Pearson changes because the transform replaces arbitrary raw spacing with a uniform rank spacing, often enhances linear alignment to human scores when raw scores are narrowly distributed. The variance analysis (Table 2, Figure 1) shows the mechanism: raw scores are tightly confined; rank–min–max expands and normalizes variance across language pairs.

For language pairs where the foundational model does not generate any relevant signal (e.g., EN–AR, EN–BHO in our runs), our method accurately preserves the lack of correlation. Our rank-based calibration is designed to normalize and rescale an existing signal; it cannot create a signal in the absence of one. Therefore, these low- or negative-correlation cases highlight the shortcomings of the underlying base model, rather than failure of the calibration process itself.

7 Reproducibility

All code, scripts, and CSV outputs used to generate the figures and tables are provided in the public repository at https://github.com/SUJAL390/rankedcomet-wmt25-emnlp.

Key files:

- `notebooks/wmt25-task1-quality prediction-sprint2.ipynb` – COMET inference.
- `scripts/calibrate_scores_ranked.py` – per-language rank–min–max calibration used to create `segments.tsv`.
- `scripts/compare_raw_ranked.py` – raw vs ranked diagnostics (produced `raw_vs_ranked_stats.csv`).
- `scripts/rankedcomet_full_analysis.py` – variance analysis, dev-based calibration recipe, ablation suite (produced variance CSVs and figures).
- `scripts/rankedcomet_figures.py` – figure production scripts.

We provide exact commands in the repository README for reproducing the full analysis.

Dev-based calibration recipe (if test-set statistics are not allowed) If a protocol forbids using test-set statistics, a held-out dev set can be used to compute a monotonic mapping (quantile interpolation or isotonic regression) from dev raw scores to quantiles, then apply that mapping to test raw scores. We include a ready-to-run script that implements this recipe in `scripts/rankedcomet_full_analysis.py`.

8 Limitations

- The 5th-place claim references a preliminary Codabench snapshot based on pseudo-gold; final human-judgment rankings may vary.
- Ranked calibration cannot produce a meaningful signal for language pairs where the foun-

dational model produces none (e.g., EN–AR, EN–BHO in our runs).

- If organizers disallow test-set statistics, apply the dev-set mapping recipe we provide. We documented this and included dev-based ablations when dev data are available.

9 Conclusion

We demonstrate that a simple per-language rank–min–max calibration applied to a robust 2022 COMET model yields a competitive QE submission on the preliminary Codabench snapshot in the WMT 2025 preliminary evaluation snapshot. The approach is affordable, deterministic, and reproducible; our diagnostics show why it improves aggregated Pearson (variance equalization) while preserving ordinal relations.

Acknowledgments

We thank the WMT organizers and reviewers for constructive feedback. Code and analysis scripts are available at https://github.com/SUJAL390/rankedcomet-wmt25-emnlp.

References

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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