UvA-MT at WMT25 Evaluation Task: LLM Uncertainty as a Proxy for Translation Quality

Di Wu **Christof Monz**
Language Technology Lab
University of Amsterdam
{d.wu, c.monz}@uva.nl

Abstract

This year, we focus exclusively on using the uncertainty quantification as a proxy for translation quality. While this has traditionally been regarded as a form of **unsupervised** quality estimation, such signals have been overlooked in the design of the current metric models—we show their value in the context of LLMs. More specifically, in contrast to conventional unsupervised QE methods, we apply recent calibration technology (Wu et al., 2025b) to adjust translation likelihoods to better align with quality signals, and we use the **single** resulting model to participate in both the general translation and QE tracks at WMT25.

Our offline experiments show some advantages: 1) uncertainty signals extracted from LLMs, like Tower or Gemma-3, provide accurate quality predictions; and 2) calibration technology further improves this QE performance, sometimes even surpassing certain metric models that were trained with human annotations, such as CometKiwi. We therefore argue that uncertainty quantification (confidence), especially from LLMs, can serve as a strong and complementary signal for the metric design, particularly when human-annotated data are lacking. However, we also identify limitations, i.e., its tendency to assign disproportionately higher scores to hypotheses generated by the model itself.

1 Introduction

In this paper, we describe the details of our submission to the WMT 2025 MT evaluation subtask-1, i.e., segment-level Quality Estimation (QE), which includes 16 translation directions. This year, we focus exclusively on using the uncertainty quantification as a proxy for translation quality. While this has traditionally been regarded as a form of unsupervised quality estimation (Fomicheva et al., 2020), such signals have been overlooked in recent designs of metric models. In this competition, we aim to examine the strengths and weaknesses of this signal.

Previous unsupervised quality estimation focused on using the model's internal information to quantify the confidence/certainty of a given translation sentence pair, e.g., using likelihood, entropy, or uncertainty signals under a Monte Carlo (MC) dropout framework (Fomicheva et al., 2020). Notably, they are relying on signals derived from the model itself and are mostly training-free.

We apply recent calibration technology (Wu et al., 2025b) for this year's competition. Unlike traditional unsupervised QE, this method aims to calibrate translation likelihood with quality during training time.

By extensive experiments, several key advantages of calibrated models can be shown as follows:

- Translation quality can be substantially improved with limited training, e.g., 2K instances for each translation direction, and the effectiveness of maximum *a posterior* decoding, like beam search, can be better realized, showing strong promise for real-world use;
- **At the same time, it provides a unified view for optimizing translation quality and estimation, which matches our goal in this competition, i.e., using the model's confidence as a proxy for translation quality.**

Our offline experiments show that the resulting model's QE ability sometimes even surpasses some accurate metric models, like cometkiwi-22 (Rei et al., 2022)[1], without using any human-annotated data. We therefore argue that uncertainty quantification, especially from LLMs, can serve as a strong and complementary signal for the metric design, particularly when human-annotated data are lacking.

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

Proceedings of the Tenth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 974–983
November 8–9, 2025.

However, our preliminary tests on this year's WMT25 test set also reveal limitations in using uncertainty signals for quality estimation—**notably, they tend to assign significantly higher scores to hypotheses generated by the model itself**—a pattern that is, to some extent, consistent with recent findings that LLM-as-a-judge systems tend to favor their own outputs (Panickssery et al., 2024).

In the next sections, we will briefly describe: 1) the framework of calibration, 2) offline experiments for both translation performance and our QE strategy by using the calibration approach, 3) the implementation for our submission at the WMT25-QE track, and 4) a discussion of the strengths and limitations of using model confidence for quality estimation.

2 Calibration Method

In this section, we briefly describe the calibration method to keep the paper self-contained; please refer to (Wu et al., 2025b) for details.

Formally, given a parameterized auto-regressive language model p_θ and a translation instruction x, the *log-likelihood* of a translation hypothesis y_i is denoted as $z_\theta(y_i|x) = \log p_\theta(y_i|x)$. Meanwhile, the quality of this translation can be defined as $q(y_i|x)$ where q represents any external quality evaluation model. When sampling hypotheses y from p_θ conditioned on x, both $z_\theta(y|x)$ and $q(y|x)$ can be viewed as random variables defined over the output space. This method is to calibrate the likelihoods of generated hypotheses with their quality to maximize the correlation between $z_\theta(y|x)$ and $q(y|x)$.

The calibration method uses the statistic, *Pearson correlation coefficient* $\rho(a, b)$, to quantify the correlation. Let $a, b : \mathcal{Y} \to \mathbb{R}$ be two real-valued functions defined over a domain $\mathcal{Y}$. The corresponding Pearson score between a and b is given by

$$\rho(a, b) = \frac{\mathbb{E}_{y\sim p}\left[(a(y) - \mu_a)\,(b(y) - \mu_b)\right]}{\sigma_a \sigma_b}, \quad (1)$$

where μ_a, μ_b and σ_a, σ_b denote expectations and standard deviations, respectively. This formulation computes the correlation by normalizing the expected product of their centered values. Due to its scale-invariance and ability to capture trend consistency, the Pearson correlation coefficient is widely used in translation metric meta-evaluation.

The calibration method calculates and optimizes ρ with respect to the likelihood of hypotheses $z_\theta(y|x)$ and the quality score $q(y|x)$. Practically, given the intractably large decoding space, it employs Monte Carlo sampling for approximation. For each source sentence x, we generate k hypotheses y_i ($i \in \{1, \ldots, k\}$) by repeatedly prompting a large language model θ with nucleus sampling, and compute the corresponding $z_\theta(y_i|x)$ and $q(y_i|x)$, and estimate the corresponding μ_z, μ_q and σ_z, σ_q. Accordingly, we define the Pearson-based loss using estimates under the nucleus-induced distribution $\tilde{p}$ as follows:

$$\mathcal{L}_{\mathrm{p}} = -\frac{1}{k} \sum_{\substack{i=1 \\ y_i\sim\tilde{p}_\theta(\cdot|x)}}^{k} \frac{z_\theta(y_i|x) - \mu_z}{\sigma_z} \cdot \frac{q(y_i|x) - \mu_q}{\sigma_q}. \quad (2)$$

It additionally introduces a supervised fine-tuning (SFT) term on the highest-scoring samples as a regularizer to ensure that the model's likelihood distribution remains grounded in high-quality translations, since the Pearson objective alone enforces correlation but does not constrain the absolute scale. The final loss for calibration is formulated as $\mathcal{L}_{\text{cal}} = \mathcal{L}_{\mathrm{p}} + \mathcal{L}_{\text{sft}}$.

An off-policy formulation can be obtained by trivially replacing the current model p_θ with an external model p_{θ^*} for sampling. Overall, by minimizing $\mathcal{L}_{\text{cal}}$, we encourage the Pearson score between z and q to increase. In practice, we use a gradient-based optimizer, Adam, to optimize θ for this goal, with gradients propagated through z_θ, μ_z, and σ_z. Despite its simplicity, several important characteristics are captured in this formulation:

- It models hypothesis qualities from a holistic view, enabling the model to make finer-grained distinctions in translation quality within the decoding space.
- It considers the value of translation quality by the metric function $q(\cdot|x)$, which is ignored in virtually all existing methods based on Bradley-Terry and Plackett-Luce, such as CPO.
- Pearson's correlation inherently applies normalization to a group of both likelihood and quality points. This normalization makes the objective invariant to scale and shift, thereby promoting stable and robust optimization across diverse input distributions.
- **The objective, i.e., the Pearson's score itself, is inherently shared with that of translation metric meta-evaluation, offering a unified perspective for both quality optimization and es-**

timation. Meanwhile, unlike other statistics like Spearman's or Kendall's scores, Pearson's coefficient is differentiable and thus suitable for gradient-based optimization frameworks.

3 Offline Evaluation

In this section, we demonstrate the effectiveness of the calibration method by applying it to the strong LLM-based translation system, i.e., Tower (Rei et al., 2024). We briefly show this method's capabilities in both (1) translation quality optimization and (2) quality estimation optimization.

Note that our online submission system is based on Gemma-3-12B (see Section 4), as it covers more languages for WMT25 than Tower.

3.1 Experimental Setups

Base model. We conduct experiments on Tower-7B, Tower-13B, and TowerMistral-7B models.

Calibration dataset. For the training set, we firstly merge all English sentences from the Flores-200 dataset (Costa-Jussà et al., 2022) in *dev* and *devtest* splits and use them as the source, consisting of 2,009 samples. We focus on off-policy experiments, where we use these sentences to construct translation prompts for each direction by calling an external strong translation model, rather than sampling from the base model itself. Here, we query `gpt-4o-mini`[2] 16 times per prompt, employing nucleus sampling with a temperature of 1.0 and a top-p of 0.98. The resulting bitexts are evaluated using CometKiwi-XXL to reflect corresponding quality scores. In this study, we only use this small-scale dataset to post-train our model. This is motivated by recent studies (Xu et al., 2023; Wu et al., 2024) showing that a few high-quality samples with a strong base model can significantly enhance the system's performance.

Training setups. For all experiments, we train models using LoRA (Hu et al., 2022) with rank 8, setting α to 32 and dropout to 0.05. Training uses a batch size of 32, gradient accumulation of 8 steps, and sequences capped at 512 tokens. To ensure robust results, we experiment with learning rates ranging from 1e-5 to 1e-4, reporting the best results for all settings. Adam (Kingma and Ba, 2014) is used as an optimizer. All experiments use H100 GPUs, with 7B models trained on one GPU and 13B models trained on two GPUs.

3.2 Translation Quality Results

Table 1 presents the results for the `Tower` series under an off-policy setting, measured by CometKiwi-XL and XCOMET. Except for closed-source models, all results are decoded by beam search with a beam size of 5. TowerInstruct-7B/-13B, and TowerInstruct-Mistral-7B are official implementations (Rei et al., 2024), supervised fine-tuned (SFT) on the corresponding base models using TowerBlock. We also conducted SFT on the Tower-Base series using 2K Best-of-N samples per direction, selected from our calibration dataset (§3.1) based on the highest CometKiwi-XXL scores. The resulting performance is comparable to the official instruction models.

When applying our calibration approach, very strong improvements can be observed across all directions, metrics, and base models. First, it leads to an average improvement of +2.8 points in KIWI-XL and +2.7 points in XCOMET over TowerInstruct-Mistral-7B. Additionally, Table 2 shows gains of +3.6 points in KIWI-XXL and +1.2 points in COMET, respectively. Second, this performance is comparable to that of the current top-performing system, that is Tower-70B-v2 equipped with 100-time-sampling MBR/TRR[3], while being approximately 200 times faster[4].

We also compare our approach with CPO (Xu et al., 2024), a widely used preference optimization method for translation. Following its original setup, we select the highest- and lowest-scoring candidates as accepted and rejected samples, respectively, and achieve consistent, substantial improvements over CPO.

3.3 Quality Estimation Results

As detailed in §2, we shared the objective for translation quality optimization and estimation, although supervisions are from machine annotations instead of human annotations. If optimized effectively, the resulting model should inherently acquire the ability to assess translation quality using

[2]Recent study (Wu et al., 2025a) shows that GPT-4o-mini can already serve as a strong translation system.

[3]TRR (Rei et al., 2024) denotes an ensemble strategy that applies reranking based on multiple metric model to select the best candidate from multiple sampled hypotheses. They report TRR results when it surpasses MBR.

[4]We roughly estimate the latency of the Tower-70B-v2 model to be 10 times that of the Tower-Mistral-7B model. Meanwhile, the former employs 100× sampling, while the latter uses beam search with a beam size of 5.

	Models	en→de KIWI-XL	en→de XCOMET	en→es KIWI-XL	en→es XCOMET	en→ru KIWI-XL	en→ru XCOMET	en→zh KIWI-XL	en→zh XCOMET	en→fr KIWI-XL	en→fr XCOMET
Closed	GPT-4o-mini	68.3	91.7	70.2	87.0	68.1	81.6	69.0	79.7	65.6	83.0
	GPT-4o	68.6	92.6	70.6	87.7	69.1	83.4	69.9	81.3	66.0	83.9
	Tower-70B-v2	–	–	–	–	–	–	–	–	–	–
	Tower-70B-v2 + MBR/TRR	72.3	–	74.5	–	74.2	–	72.6	–	–	–
	TowerInstruct-7B	69.0	91.7	70.8	86.9	69.0	81.5	68.5	78.7	67.9	84.1
	TowerBase-7B	–	–	–	–	–	–	–	–	–	–
	+ SFT on BoN data	70.0	92.0	70.8	86.5	69.6	81.6	68.4	77.9	68.0	83.7
	+ CPO	71.1	93.1	72.0	87.6	71.6	83.8	70.4	80.9	69.3	85.8
	+ Calibration (ours)	**71.6**	**93.6**	**73.5**	**89.0**	**72.4**	**84.8**	**70.4**	**81.0**	**70.0**	**86.8**
	TowerInstruct-13B	69.9	92.5	71.8	87.7	70.6	83.3	70.1	80.8	68.1	85.1
	TowerBase-13B	–	–	–	–	–	–	–	–	–	–
	+ SFT on BoN data	71.1	92.7	71.8	87.5	71.3	82.8	70.1	80.0	68.0	84.4
	+ CPO	70.5	92.2	72.0	87.7	71.9	84.0	70.3	81.4	68.8	85.5
	+ Calibration (ours)	**72.5**	**94.2**	**73.8**	**90.0**	**73.6**	**86.4**	**72.1**	**83.6**	**70.8**	**87.5**
	TowerInstruct-Mistral-7B	70.0	92.6	71.9	87.5	70.3	83.3	69.6	80.4	68.3	84.7
	+ SFT on BoN data	70.7	92.7	71.8	87.1	70.8	82.9	70.5	80.4	68.5	84.4
	+ CPO	71.2	93.0	73.1	89.0	72.3	85.1	71.8	83.6	70.0	86.9
	+ Calibration (ours)	**72.4**	**94.0**	**73.9**	**89.9**	**73.6**	**86.1**	**72.6**	**83.7**	**70.8**	**87.4**

	Models	en→nl KIWI-XL	en→nl XCOMET	en→it KIWI-XL	en→it XCOMET	en→pt KIWI-XL	en→pt XCOMET	en→ko KIWI-XL	en→ko XCOMET	Avg. KIWI-XL	Avg. XCOMET
Closed	GPT-4o-mini	69.4	88.9	68.1	83.7	71.2	87.6	73.2	84.2	69.2	85.3
	GPT-4o	70.6	90.5	68.7	85.7	71.5	88.5	73.7	85.6	69.8	86.6
	Tower-70B-v2	–	–	–	–	–	–	–	–	–	–
	Tower-70B-v2 + MBR/TRR	–	–	–	–	–	–	–	–	–	–
	TowerInstruct-7B	71.5	90.9	71.1	86.1	71.1	86.8	73.6	82.8	70.3	85.5
	TowerBase-7B	–	–	–	–	–	–	–	–	–	–
	+ SFT on BoN data	71.5	89.6	70.8	85.4	72.5	87.6	75.7	84.1	70.8	85.4
	+ CPO	71.9	90.9	72.2	86.7	73.4	88.7	76.1	87.2	72.0	87.2
	+ Calibration (ours)	**73.3**	**91.9**	**73.5**	**88.1**	**74.8**	**89.9**	**76.8**	**87.2**	**72.9**	**88.0**
	TowerInstruct-13B	71.7	91.0	71.1	87.3	72.1	88.2	75.4	84.8	71.2	86.7
	TowerBase-13B	–	–	–	–	–	–	–	–	–	–
	+ SFT on BoN data	71.7	90.4	71.6	86.1	73.0	88.1	76.2	85.2	71.6	86.4
	+ CPO	72.3	90.8	72.5	87.4	72.2	86.9	76.9	87.9	71.9	87.1
	+ Calibration (ours)	**73.9**	**92.6**	**73.9**	**89.3**	**75.2**	**90.4**	**78.0**	**89.5**	**73.8**	**89.3**
	TowerInstruct-Mistral-7B	71.9	91.1	71.6	87.2	72.1	88.0	74.2	85.6	71.1	86.7
	+ SFT on BoN data	72.3	90.7	71.6	86.2	72.7	87.9	76.2	86.0	71.7	86.5
	+ CPO	73.3	92.3	73.1	88.5	74.0	89.7	77.4	89.3	72.9	88.6
	+ Calibration (ours)	**74.2**	**93.2**	**74.1**	**89.6**	**75.1**	**90.7**	**78.1**	**89.7**	**73.9**	**89.4**

Table 1: en→xx translation qualities on WMT24 measured by CometKiwi-XL and XCOMET. Note that the Tower-v2 models, including Tower-70B-v2, have not been publicly released. We report their best results as published by Rei et al. (2024). For GPT-4o and GPT-4o-mini, we use the prompts following (Hendy et al., 2023). Results in other metrics can be found in Appendix A. Notably, according to Kocmi et al. (2024), improvements of ≥ 1.99 in XCOMET or ≥ 0.94 in COMET scores correspond to at least 90% estimated accuracy in human judgment—both of which are achieved by our method.

the hypothesis *log-likelihood* as a metric. In this section, we evaluate how effectively calibration can elicit this capability.

We use the WMT22 metric meta-evaluation dataset (Zerva et al., 2022) and follow the official practice to assess quality estimation ability using Spearman's and Kendall's correlation. We evaluate all training directions on Tower that overlap with the WMT dataset, namely, en→de and en→ru.

Figure 1 depicts the Spearman score (metric performance) and the corresponding translation performance under different settings for Tower-7B and Tower-13B, including: (1) supervised fine-tuning using varying amounts of best-of-N samples (400/800/1200/1600/2000 samples per direction), (2) scaling the base model size from 7B to 13B, and (3) applying our calibration method. It shows that:

(1) As more Best-of-N samples are included in SFT, translation performance progressively improves. Interestingly, the quality estimation ability (Spearman scores) increases from around 51.5 to 54.0 points. We attribute this to the fact that the model assigns higher likelihoods to better hypotheses. However, these improvements are limited and not general across languages, see Appendix B.

(2) Examining the effects of scaling, we observe that: (i) scaling up from 7B to 13B generally improves translation performance for both the original TowerInstruct models and the fine-tuned models; (ii) however, its impact on calibration, i.e., quality estimation ability, remains minimal.

(3) Our calibration method manifests very strong improvements in both translation and quality estimation. For example, when applying our method to TowerBase-13B, the resulting model surpasses some state-of-the-art systems in both translation performance and quality estimation ability, i.e., Tower-70B-v2+MBR/TRR and CometKiwi, at the same time.

Similar trends can be found in Figure 2 when we use the Kendall coefficient to measure the correlation. Results for en→ru are provided in Appendix B.

Overall, we observe a clear, albeit sometimes non-linear, correlation between the models' translation performance and their quality estimation ability. These results suggest—to some extent—a uni-

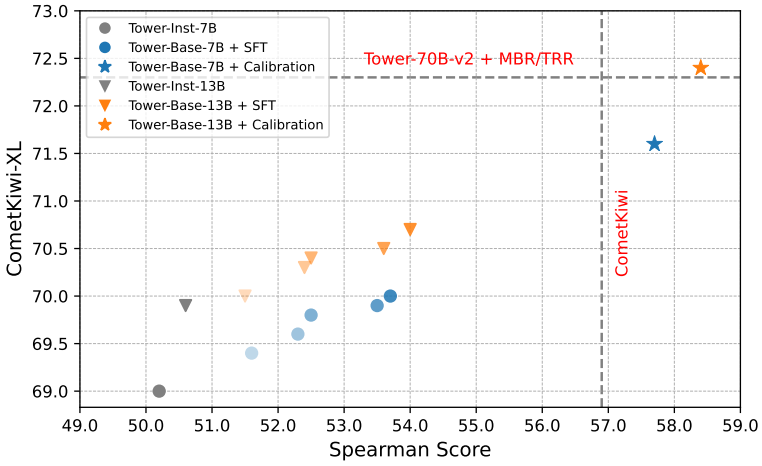

Figure 1: The Spearman coefficient and the corresponding translation performance in en→de direction under different settings for the Tower series models. The color gradients of ▼ and ●, from lighter to darker shades, indicate the results of fine-tuning with varying amounts of Best-of-N data, from 400 to 2000 samples. ★ denotes the application of our calibration method, which simultaneously surpasses both the state-of-the-art translation system and the widely used quality estimation model.

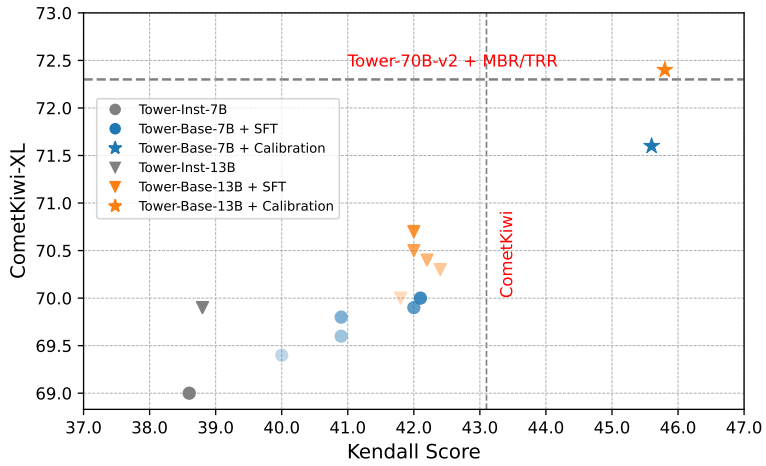

Figure 2: The Kendall coefficient and the corresponding translation performance in en→de direction under different settings for the Tower series models. The color gradients of ▼ and ●, from lighter to darker shades, indicate the results of fine-tuning with varying amounts of Best-of-N data, from 400 to 2000 samples. ★ and ★ denote the application of our calibration method on 13B and 7B models, respectively.

fied perspective: a well-performing translation system should inherently 'know' what constitutes a good translation. In turn, we also suggest optimizing translation quality by improving calibration on LLMs, rather than relying solely on extreme scaling or supervised fine-tuning, as the latter approaches show relatively limited effectiveness.

4 Implementation of Our WMT25 Submission

In this section, we describe the implementation of our QE system for WMT25. To cover more languages, we use Gemma-3-12B as the base model instead of Tower, as noted in Section 3.1.

To construct the calibration dataset, similar to that in Section 3.1, we feed the source segments in the WMT25 general translation test set into GPT-4o-mini, using the prompts provided with the official test set[5], to generate 16 hypotheses per sample. Note that we use the WMT25 blind test sets rather than Flores. This choice is motivated by (1) better alignment with the domain used in testing and (2) consistency with WMT25's paragraph-level data format.

Following that in Section 3.1, the corresponding hypotheses are decoded using nucleus sampling with a top-p of 0.98 and a temperature of 1.0. Each sample is at the paragraph level, where "$\backslash n$" re-

[5]Official prompts can be found here.

mains in the original data as a separator. We also use CometKiwi-XXL to score each one-to-many translation pair in our synthetic dataset.

We apply the calibration method (Wu et al., 2025b), as we mentioned in Section 2, as the only post-training method on Gemma-3-12B.

Finally, the resulting model, trained on synthetic data derived from the WMT25 general translation test set, is used as a quality estimation model here. For each sample in the WMT25 QE dataset, we feed the source and target segments into our model (with the corresponding prompt) and directly use the **average log-likelihood** of the target segments as the quality assessment scores for submission.

For the performance of our system on the WMT25-QE track, please refer to this year's findings paper, which has not yet been officially released at the time of this submission.

5 On the Limitation of Using Uncertainty as a Proxy for Translation Quality

Although we demonstrate the effectiveness of using LLM uncertainty as a proxy for translation quality in Section 3.3, we also identify an important limitation—**this method will give significantly higher scores for translations that are from itself**—it favors its own output when it uses maximum a *posteriori* as the decoding rule.

This issue was not identified by (Wu et al., 2025b), as their QE testing set lacks such special conditions and does not include the model's own translation outputs during QE evaluation.

Meanwhile, we argue that this issue is likely to be a general limitation of most metrics based on translation uncertainty. More broadly, it to some extent aligns with observations about LLM-as-a-judge (Panickssery et al., 2024), where LLMs tend to favor their own generations.

Lastly, we anticipate that the official WMT25 QE test set will be particularly challenging for our metric. **As noted above, we use a single calibrated model to participate in both the general translation and quality estimation tasks at WMT25.** Therefore, if any hypotheses generated by our system appear in this year's QE test set, it is likely to assign them the highest scores. We have found indications of this, although we cannot confirm it because the official description of the test set has not yet been released.

References

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

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.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Tom Kocmi, Vilém Zouhar, Christian Federmann, and Matt Post. 2024. Navigating the metrics maze: Reconciling score magnitudes and accuracies. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1999–2014, Bangkok, Thailand. Association for Computational Linguistics.

Arjun Panickssery, Samuel Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems*, 37:68772–68802.

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, Jose Pombal, Nuno M. Guerreiro, João Alves, Pedro Henrique Martins, Patrick Fernandes, Helena Wu, Tania Vaz, Duarte Alves, Amin Farajian, Sweta Agrawal, Antonio Farinhas, José G. C. De Souza, and André Martins. 2024. Tower v2: Unbabel-IST 2024 submission for the general MT shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 185–204, Miami, Florida, USA. Association for Computational Linguistics.

Di Wu, Seth Aycock, and Christof Monz. 2025a. Please translate again: Two simple experiments on whether human-like reasoning helps translation. *arXiv preprint arXiv:2506.04521*.

Di Wu, Yibin Lei, and Christof Monz. 2025b. Calibrating translation decoding with quality estimation on llms. *arXiv preprint arXiv:2504.19044*.

Di Wu, Shaomu Tan, Yan Meng, David Stap, and Christof Monz. 2024. How far can 100 samples go? unlocking zero-shot translation with tiny multi-parallel data. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15092–15108, Bangkok, Thailand. Association for Computational Linguistics.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. A paradigm shift in machine translation: Boosting translation performance of large language models. *arXiv preprint arXiv:2309.11674*.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *arXiv preprint arXiv:2401.08417*.

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.

A Results based on Tower in KIWI-XXL and COMET

Table 2 shows off-policy results measured by two other metrics, i.e., CometKiwi-XXL (abbreviated as KIWI-XXL) and COMET-22 (abbreviated as COMET). Very strong average performance improvements can be observed. For instance, +3.6 and +1.1 points of KIWI-XXL and COMET average gains are shown over TowerInstruct-Mistral-7B.

B Results on En→Ru Direction

In this appendix, we provide additional results in en→ru, complementing Section 3.3 of the main text.

Figure 3 shows the Spearman coefficient and the corresponding translation performance in the en→ru direction. Meanwhile, Figure 4 present the results using Kendall's τ for en→ru direction. It is clear that the main findings, as mentioned in Section 3.3, hold across language directions and statistics.

	Models	en→de		en→es		en→ru		en→zh		en→fr	
		KIWI-XXL	COMET	KIWI-XXL	COMET	KIWI-XXL	COMET	KIWI-XXL	COMET	KIWI-XXL	COMET
Closed	GPT-4o-mini	76.4	82.7	76.3	83.8	75.5	82.5	75.8	84.6	74.7	81.5
Closed	GPT-4o	77.7	82.5	77.3	83.8	77.6	82.8	77.6	84.5	76.2	81.7
Closed	Tower-70B-v2	–	–	–	–	–	–	–	–	–	–
Closed	Tower-70B-v2 + MBR/TRR	–	–	–	–	–	–	–	–	–	–
	TowerInstruct-7B	76.5	81.2	76.3	82.8	75.9	81.1	74.8	83.1	76.7	81.2
	TowerBase-7B	–	–	–	–	–	–	–	–	–	–
	+ SFT on BoN data	77.2	81.3	75.8	82.4	76.2	80.9	74.4	82.4	76.2	81.0
	+ CPO	78.9	82.2	78.0	83.2	78.8	82.2	77.8	83.4	78.7	81.2
	+ Calibration	**79.5**	**82.8**	**79.8**	**83.7**	**80.4**	**82.9**	**78.0**	**83.2**	**80.2**	**81.7**
	TowerInstruct-13B	78.1	82.3	77.6	83.5	78.2	82.1	76.9	83.8	77.4	81.6
	TowerBase-13B	–	–	–	–	–	–	–	–	–	–
	+ SFT on BoN data	79.0	82.3	77.0	83.1	78.4	82.0	76.8	83.8	77.2	81.5
	+ CPO	79.1	82.1	78.6	82.5	80.3	82.6	78.0	83.4	79.3	81.5
	+ Calibration	**81.3**	**83.4**	**80.9**	**84.1**	**82.3**	**83.8**	**80.4**	**84.5**	**81.5**	**82.2**
	TowerInstruct-Mistral-7B	78.1	82.0	77.9	83.0	77.9	81.8	76.6	83.8	77.6	81.5
	+ SFT on BoN data	78.3	82.0	77.5	82.9	78.3	81.5	77.3	84.0	77.3	81.4
	+ CPO	79.6	82.2	79.9	83.3	80.5	82.7	79.7	84.8	79.9	81.8
	+ Calibration	**80.7**	**83.1**	**80.6**	**83.9**	**82.0**	**83.6**	**80.4**	**84.9**	**80.8**	**82.1**

	Models	en→nl		en→it		en→pt		en→ko		Avg.	
		KIWI-XXL	COMET	KIWI-XXL	COMET	KIWI-XXL	COMET	KIWI-XXL	COMET	KIWI-XXL	COMET
Closed	GPT-4o-mini	78.3	84.6	74.1	83.6	77.9	81.9	81.2	86.2	76.7	83.5
Closed	GPT-4o	80.7	84.6	76.0	83.8	79.1	81.9	82.3	86.2	78.3	83.5
Closed	Tower-70B-v2	–	–	–	–	–	–	–	–	–	–
Closed	Tower-70B-v2 + MBR/TRR	–	–	–	–	–	–	–	–	–	–
	TowerInstruct-7B	81.1	84.4	77.7	83.7	77.9	81.8	80.0	84.7	77.4	82.7
	TowerBase-7B	–	–	–	–	–	–	–	–	–	–
	+ SFT on BoN data	80.5	83.5	76.9	83.4	78.6	81.5	82.3	85.3	77.6	82.4
	+ CPO	81.9	83.8	79.4	83.7	80.5	81.8	83.7	85.8	79.7	83.0
	+ Calibration	**83.6**	**84.8**	**81.0**	**84.3**	**81.9**	**82.7**	**84.6**	**86.1**	**81.0**	**83.6**
	TowerInstruct-13B	81.4	84.6	78.4	84.2	79.1	82.5	82.9	85.5	78.9	83.4
	TowerBase-13B	–	–	–	–	–	–	–	–	–	–
	+ SFT on BoN data	80.8	84.3	77.9	83.8	79.5	81.7	83.6	85.7	78.9	83.1
	+ CPO	82.5	84.2	80.2	83.8	79.2	80.7	85.0	86.5	80.2	83.0
	+ Calibration	**84.5**	**85.1**	**82.1**	**84.6**	**82.8**	**82.7**	**86.2**	**87.1**	**82.4**	**84.2**
	TowerInstruct-Mistral-7B	81.5	84.6	79.0	84.0	79.3	82.2	81.7	85.3	78.8	83.1
	+ SFT on BoN data	81.4	84.2	78.4	83.7	79.6	81.7	83.8	86.1	79.1	83.0
	+ CPO	83.9	84.8	80.7	84.0	81.9	82.2	85.9	86.9	81.3	83.6
	+ Calibration	**84.6**	**85.2**	**82.3**	**84.8**	**83.2**	**83.0**	**86.9**	**87.3**	**82.4**	**84.2**

Table 2: Evaluation of en→xx translation on WMT24 using CometKiwi-XXL and COMET. Results are reported for all languages covered during Tower-v1 pretraining. Note that the Tower-v2 models, including Tower-70B-v2, have not been publicly released. For GPT-4o and GPT-4o-mini, we use the prompts following (Hendy et al., 2023). Notably, according to Kocmi et al. (2024), improvements of ≥ 1.99 in XCOMET or ≥ 0.94 in COMET scores correspond to at least 90% estimated accuracy in human judgment — both of which are achieved by our method.

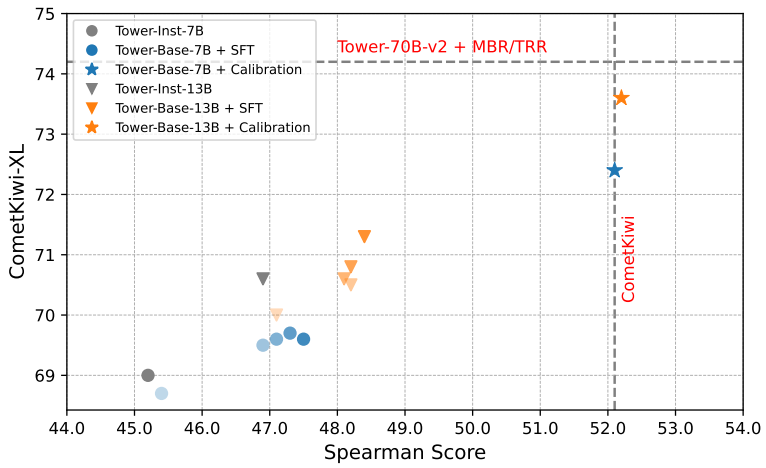

Figure 3: The Spearman coefficient and the corresponding translation performance in en→ru direction under different settings for the Tower series models. The color gradients of ▼ and ●, from lighter to darker shades, indicate the results of fine-tuning with varying amounts of Best-of-N data, from 400 to 2000 samples. ★ and ★ denote the application of our calibration method on 13B and 7B models, respectively.

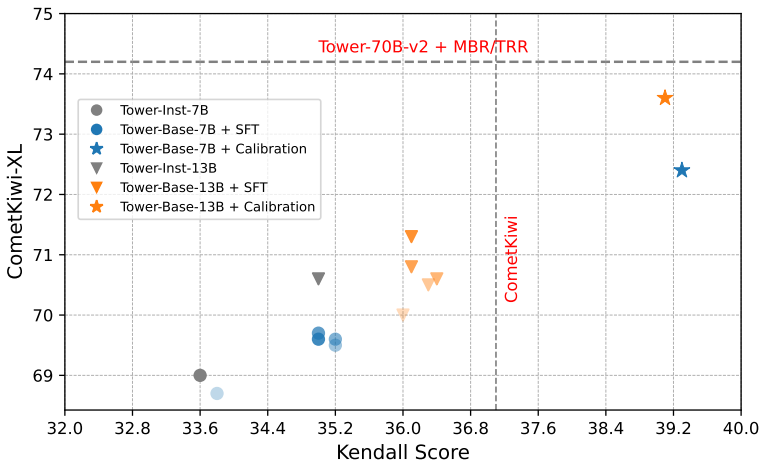

Figure 4: The Kendall coefficient and the corresponding translation performance in en→ru direction under different settings for the Tower series models. The color gradients of ▼ and ●, from lighter to darker shades, indicate the results of fine-tuning with varying amounts of Best-of-N data, from 400 to 2000 samples. ★ and ★ denote the application of our calibration method on 13B and 7B models, respectively.