Test-time Adaptation for Machine Translation Evaluation by Uncertainty Minimization

Runzhe Zhan[1], Xuebo Liu[2], Derek F. Wong[1], Cuilian Zhang[1],
Lidia S. Chao[1] and Min Zhang[2]
[1]NLP^{2}CT Lab, Department of Computer and Information Science, University of Macau
[2]Institute of Computing and Intelligence, Harbin Institute of Technology (Shenzhen)

Out-of-distribution Challenges

- **Problem:** Neural metrics were trained on News rating data.
- **Potential Risk:** Neural metrics may have robustness problems when evaluating the out-of-distribution (OOD) text.

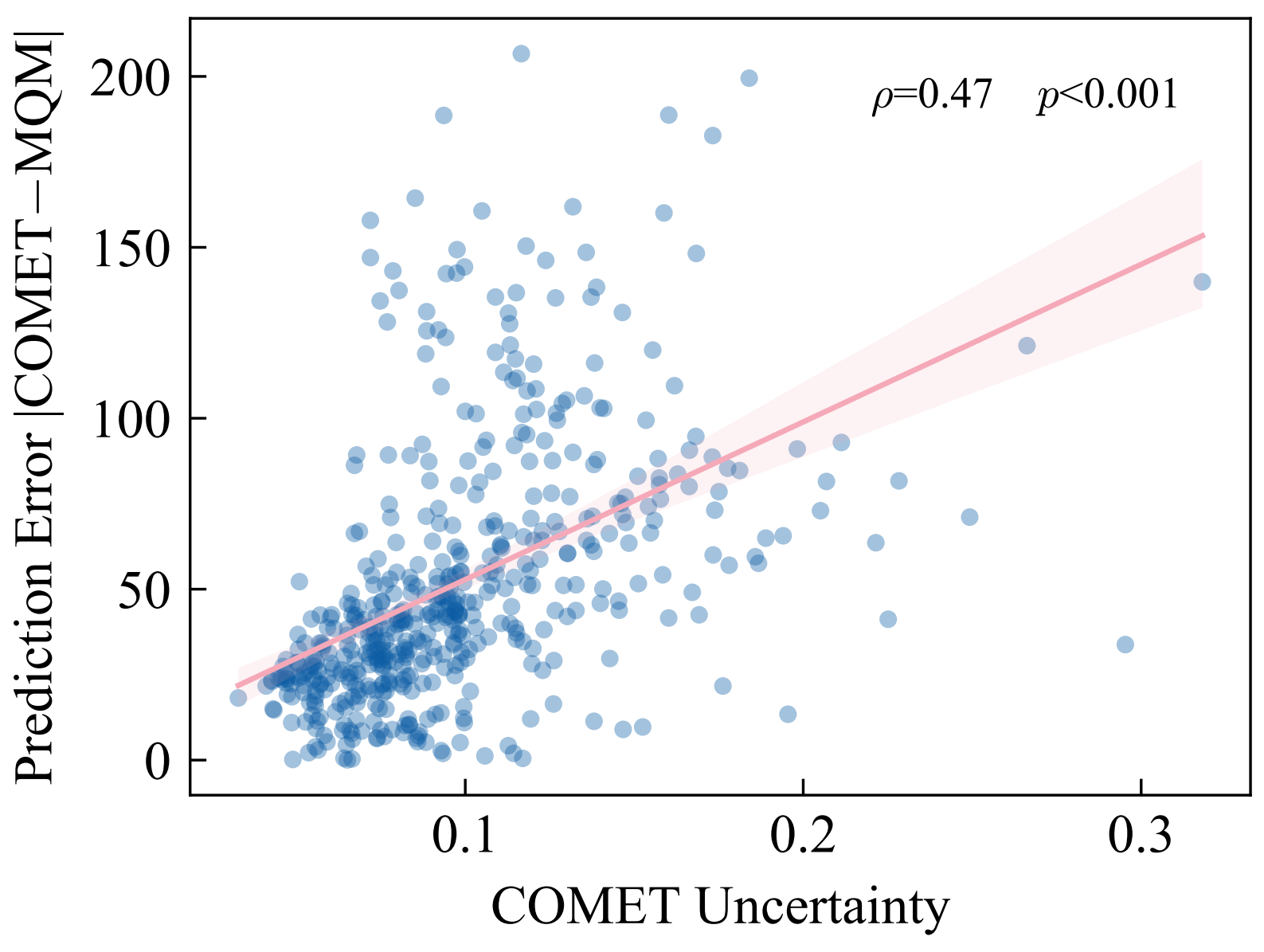
✗ **Dilemma:** Collecting multi-domain annotation data is expensive.

❖ **Main Research Goal:**

Can we alleviate OOD problem without annotated data?

Why Uncertainty Minimization?

- **Epistemic uncertainty** reflects the risk of model's predictions.
- **Observation:** Model's uncertainty positively correlates with its prediction errors. Also observed by Glushkova et al. (2021).



- **COMET:** A neural metric.
- **MQM:** Human scores.
- **Prediction Error:** Absolute Differences between metric scores and human scores.

✓ **Motivation:**

Minimize the uncertainty ➟ Minimize the prediction errors

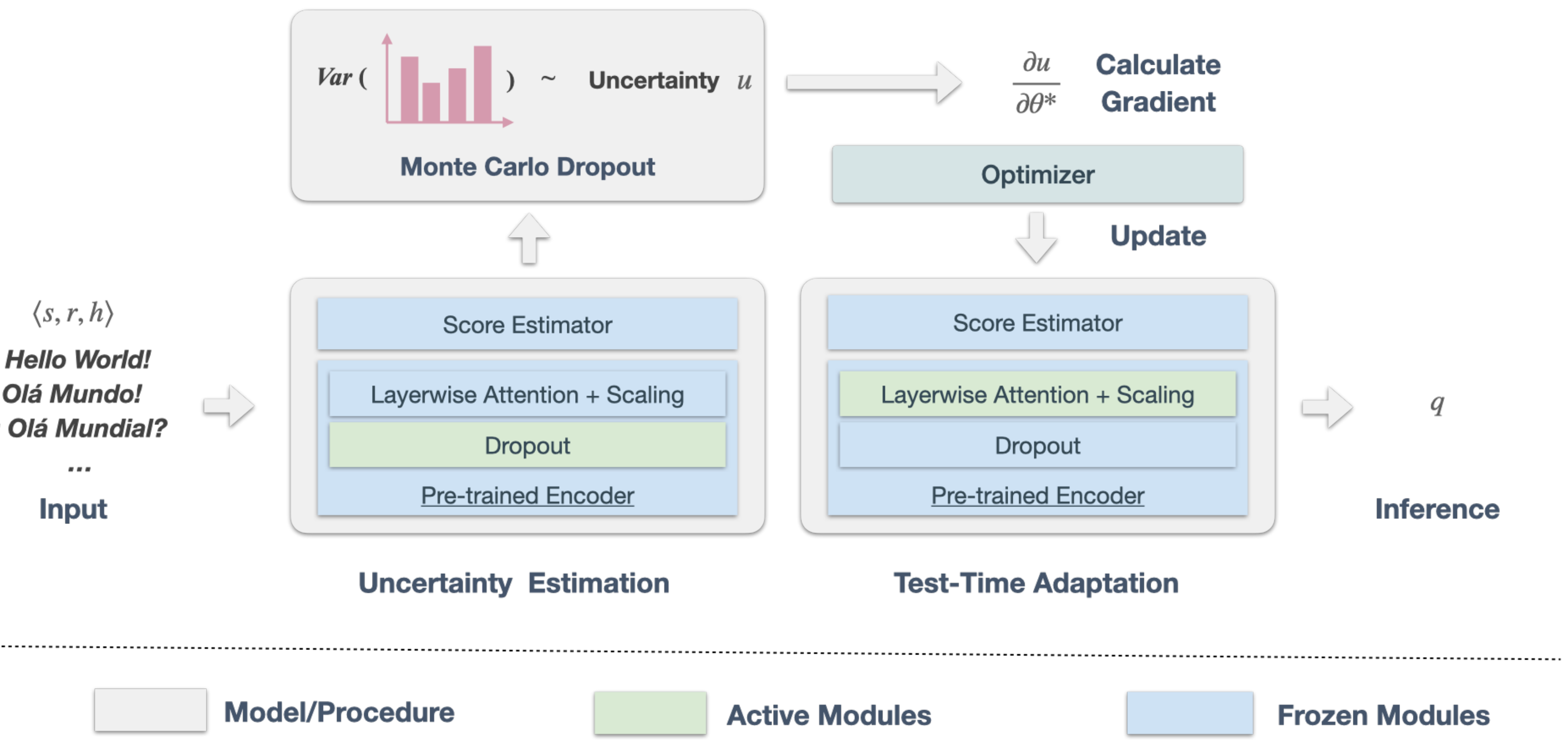
Our Proposal: TaU

- **T**est-time **A**daptation by **U**ncertainty Minimization (TaU).

✓ **Key Idea:** Make the model correct the predictions by itself through reducing the uncertainty.

✓ **Key Research Questions:**

1) How can we **estimate** the uncertainty for metrics' model?
2) How can we **reduce** the uncertainty by test-time adaptation?



TaU

- **Uncertainty Estimation:**
  - Use **Monte-Carlo Dropout** (Gal et. al, 2016; Glushkova et. al, 2021) method to estimate the uncertainty during inference.
  - Uncertainty = **Var**iance of K-times prediction

$$\underset{\substack{\uparrow\\\text{Input Data}}}{u(\langle h, s, \cdot\rangle)} = \mathbf{Var}(\{\underset{\substack{\uparrow\\\text{Metric Model (w/ Dropout)}}}{\mathrm{M}(\langle h, s, \cdot\rangle; \theta_k)}\}_{k=1}^{K})$$

- **Test-time Adaptation:**
  - Objective function: **minimize the uncertainty**
  - Do not deviate far from original parameters! Only optimize partial parameters (Layerwise Attention + Scaling Factor).

$$\theta^* = \underset{\substack{\theta^*\\\uparrow\\\text{Optimization of partial modules}}}{\arg\min}\ \mathbb{E}_{\langle h,s,\cdot\rangle \in \mathcal{D}}\ [u(\langle h, s, \cdot\rangle)]$$

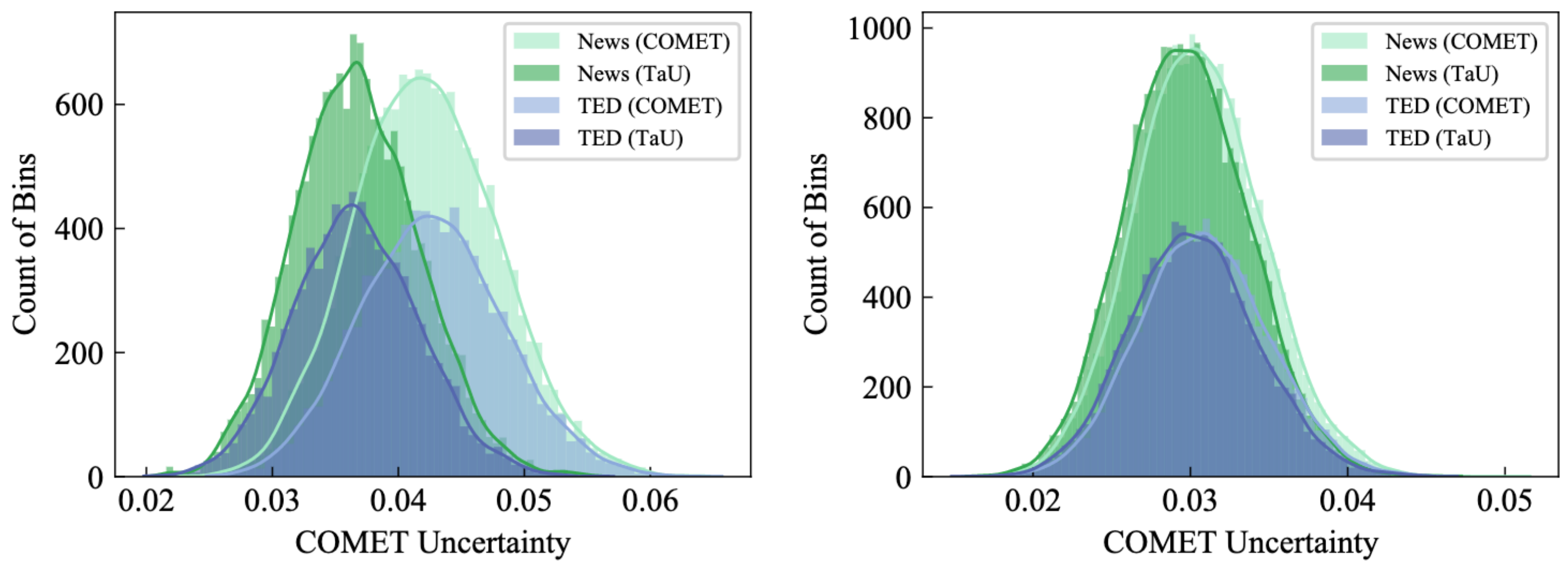
How Does TaU Work?

- Improved **system-level Pearson's correlation** performance on WMT21 MQM multi-domain benchmark.

Metrics	News w/o HT			News w/ HT			TED			Avg.
	En-De	Zh-En	En-Ru	En-De	Zh-En	En-Ru	En-De	Zh-En	En-Ru	
Baselines										
TER	93.0	41.6	-4.1	7.4	-8.5	-28.9	50.6	42.1	69.7	29.2
BLEU	93.7	31.0	50.7	13.2	-15.2	-4.3	62.0	32.4	82.8	38.5
CHRF	89.8	30.2	78.3	1.7	-14.3	12.3	47.1	36.3	82.5	40.4
BERTSCORE	93.0	54.2	62.9	7.4	9.5	-12.3	50.6	30.6	83.1	42.1
COMET-DA$_{2020}$	81.4	51.1	67.6	65.8	22.1	55.6	78.8	25.1	85.9	59.3
COMET-MQM-QE$_{2021}$	71.1	52.9	63.2	79.2	61.9	68.1	69.4	-20.9	88.4	59.3
COMET-MQM$_{2021}$	77.1	62.8	65.9	72.0	33.6	68.5	81.8	26.6	84.1	63.6
Reproduced Results and Our Methods										
◇ COMET-DA$_{2020}$	81.5	51.1	67.5	58.0	26.4	56.8	78.8	25.0	85.9	59.0
+TAU	**85.7**	**53.5**	**71.0**	48.0	**27.4**	54.5	**85.9**	**28.3**	**87.3**	**60.2**
◇ COMET-MQM-QE$_{2021}$	71.2	53.0	68.8	79.2	61.9	68.1	69.4	-20.8	81.7	59.2
+TAU	62.8	**57.4**	**70.3**	72.0	**65.2**	**78.1**	**82.9**	**25.7**	80.7	**66.1**
◇ COMET-MQM$_{2021}$	77.2	62.8	65.9	69.8	48.7	69.7	81.8	26.6	84.1	65.2
+TAU	76.5	**69.2**	**67.2**	**75.4**	**67.8**	**71.5**	**87.5**	24.5	**84.9**	**69.4**

Why Does TaU Work?

- **Validity:** Reduced the uncertainty of OOD samples.



(a) English-German (COMET-DA$_{2020}$)

(b) Chinese-English (COMET-MQM$_{2021}$)

- **Future work:** 1) Explore segment-level TaU for diverse data.
  2) Apply test-time adaptation method to LLM.

Acknowledgement

This work was supported in part by the Science and Technology Development Fund, Macau SAR (Grant Nos. FDCT/060/2022/AFJ, FDCT/0070/2022/AMJ), the National Natural Science Foundation of China (Grant No. 62206076), the Research Program of Guangdong Province (Grant No. 2220004002576), Shenzhen College Stability Support Plan (Grant Nos. GXWD20220811173340003, GXWD20220817123150002), Shenzhen Science and Technology Program (Grant No. RCBS20221008093121053) and the Multi-year Research Grant from the University of Macau (Grant No. MYRG2020-00054-FST). This work was performed in part at SICC which is supported by SKL-IOTSC, and HPCC supported by ICTO of the University of Macau. We would like to thank the anonymous reviewers and meta-reviewer for their insightful suggestions.

* The 61st Annual Meeting of the Association for Computational Linguistics (ACL), 2023.