Machine Translation Evaluation English-Thai MQM Ranking Dataset

Phichet Phuangrot[1*], **Natdanai Trintawat**[1*],
Kanawat Vilasri[1], **Yanapat Patcharawiwatpong**[1],
Pachara Boonsarngsuk[1,2], **Nat Pavasant**[1†], **Ekapol Chuangsuwanich**[1]
[1]Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University
[2]Department of Mathematics, King's College London
phichetphuangrot@gmail.com, natdanaitrintawat@gmail.com,
kanawatgrill2002@gmail.com, tsunnoreply@gmail.com
pacharawinboon@gmail.com, Nat.p@chula.ac.th, ekapolc@cp.eng.chula.ac.th

Abstract

We introduce **MEET-MR** (Machine Translation English–Thai MQM and Ranking Dataset), a comprehensive benchmark for evaluating English–Thai machine translation systems. The dataset is constructed using the Multidimensional Quality Metrics (MQM) annotation framework, providing fine-grained human judgements of translation quality. In addition, MEET-MR includes human preference rankings and reference translations, enabling both absolute and relative assessment of translation quality. The dataset covers nine diverse domains providing linguistic and contextual diversity. By combining high-quality reference translations, objective MQM error annotations, and subjective preference rankings, MEET-MR serves as a valuable resource for studying translation quality estimation, model alignment with human evaluation, and cross-domain performance in English–Thai machine translation. MEET-MR is publicly available at https://huggingface.co/datasets/Chula-AI/MEET-MR

1 Introduction

Machine translation (MT) evaluation has achieved major progress with traditional automatic metrics, such as BLEU (Papineni et al., 2002), yet these often fail to capture adequacy and fluency nuances, especially for low-resource or typologically distant languages. More recent neural models, such as BERTScore (Zhang* et al., 2020) or BLEURT (Sellam et al., 2020), have been developed to address these limitations, but reliable and consistent evaluation remains a persistent challenge.

The Workshop on Statistical Machine Translation (WMT) has standardized machine translation evaluation through large-scale human annotation campaigns. Early editions employed Relative Ranking (RR) (Callison-Burch et al., 2008), where human annotators compared multiple translations of the same source segment to determine relative quality. Later WMT evaluations introduced Direct Assessment (DA) (Graham et al., 2013, 2015), in which raters assign direct scores to individual translations. More recently, the Multidimensional Quality Metrics (MQM) framework (Lommel et al., 2014) has been used in Freitag et al. (2021a,b) and later shared tasks, offering a detailed, taxonomy-based approach to error annotation that remains the current standard in WMT evaluations.

Together, RR, DA, and MQM provide complementary evaluation signals, supporting error diagnosis, system-level comparisons, and the development of quality estimation (QE) models. However, these resources have focused primarily on high-resource language pairs such as English–German, English–Chinese, and English–Russian (Freitag et al., 2021a).

For English–Thai, existing corpora such as SCB-MT-EN-TH-2020 (Lowphansirikul et al., 2021), ALT (Thu et al., 2016), OPUS (Tiedemann, 2012), and FLORES-101/200 (Goyal et al., 2022) provide useful parallel data for translation; however, they typically include only machine translation or human translation and lack human quality evaluation under frameworks such as DA, MQM, or RR. Therefore, a gap remains in resources for systematic comparison and quality estimation, despite the growing demand for English–Thai machine translation.

We addressed this gap by introducing the **MEET-MR** dataset with human evaluation. All annotations are conducted by professional annotators using strict guidelines. The dataset provides three types of annotations: **Translations, MQM annotations, and Relative Ranking** by combining error annotations with preference rankings. Our dataset provides fine-grained annotations and preference judgments, enabling progress in MT evaluation,

[*]These authors contributed equally.
[†]Corresponding author.

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 2: Short Papers, pages 572–587

quality estimation for the English–Thai language pair.

2 MEET-MR

The **MEET-MR** dataset was developed to advance English–Thai machine translation evaluation through diverse and human-annotated data. It contains 2,142 English source sentences collected from multiple public and manually curated sources across various domains. Each sentence was translated by 10 different MT systems and paired with professional Thai reference translations. Human annotators evaluated translations using MQM framework (minor, major, and critical errors) and provided Relative Ranking reflecting overall **human preference** among translations. Compared with existing resources such as WMT MQM (Freitag et al., 2021a), MLQE-PE (Specia et al., 2022), and SCB-MT-EnTh (Lowphansirikul et al., 2021), MEET-MR offers broader domain coverage and uniquely integrates fine-grained MQM annotations with human preference rankings for comprehensive MT evaluation, as shown in Table 1. An overview of the dataset construction process is shown in Figure 1.

2.1 Sentence Curation

To create a high-quality human evaluation dataset for MT, we curated English source sentences that span a diverse range of topics and linguistic phenomena, ensuring broad coverage and reliable assessment across translation systems. The tone of our sentences ranged from written language to casual conversation. The vocabulary ranged from common everyday usage to specialized medical terminology. As a result, 2,142 unique English sentences were curated from diverse sources, covering 9 domains to provide broad vocabulary representation. Table 2 summarizes the distribution of source sentences. The source sentences are short, with an average of 18 words, following the distribution observed in the WMT MQM 2020–2021 dataset. Figure 2 shows the distribution of segment lengths. Further details on the data sources and curation process are provided in Appendix A.

2.2 Candidate Translation Generation

To ensure a rich and diverse set of translations for each source sentence, we generated candidate translations using 10 different machine translation (MT) systems, comprising both large language models (LLMs) and conventional MT systems. This diversity allows for a broad range of translation quality levels, which is essential for meaningful human evaluation. We used a total of 10 machine translation (MT) systems to generate candidate translations for each English source sentence, comprising both large language models (LLMs) (both generic LLMs and Thai-specific LLMs) and conventional MT systems. This setup ensures a wide range of model sources to represent different translation quality levels for comprehensive human evaluation.

The following 8 LLMs were employed: GPT-4o-mini[1], Claude 3.5 Sonnet[2], LLaMa3-8B-WangchanX-sft-Full (Phatthiyaphaibun et al., 2024), Typhoon-v1.5x-70B-Instruct (Pipatanakul et al., 2023), Qwen2.5-72B-Instruct (Yang et al., 2024), Grok-beta, Gemma2-9B-cpt-Sea-Lionv3-Instruct (Ng et al., 2025), and OpenThaiGPT-1.5-72B-Instruct (Yuenyong et al., 2025).

For all LLM-based systems, translations were generated using the following prompt:

```
Translate the following text to
Thai.
<input text>
```

In addition, two non-LLMs systems, Google Translate API and NLLB-200-1.3B (Koishekenov et al., 2023), were also used. For both, translations were generated by setting the source language to English and the target language to Thai via their respective configuration options. A total of 20,100 segments were collected in this manner. Certain systems such as Grok-beta were not available throughout the process of the study, and thus were not fully translated.

2.3 Annotation

In order to ensure high-quality data labeling, we designed a structured annotation process that combined qualified annotators, a custom-built annotation interface, and annotation guideline. We recruited two annotators, each holding a bachelor's degree in linguistics, to leverage their expertise in language analysis. Both annotators were compensated fairly according to local rates. As a result of this annotation process, Table 3 summarizes the distribution of MQM error types in our dataset. See details on the annotation user interfaces and annotation guideline in Appendix B.

[1] gpt-4o-mini-2024-07-18
[2] claude-3-5-sonnet-20240620

Dataset	Languages	MT Evaluation	Translation	#Domains	#Source sentences	#Total segments
WMT MQM 2021	En→De, En→Ru, Zh→En	DA, MQM, SQM	Yes	2	3,274	50,162
WMT MQM 2022	En→De, En→Ru, Zh→En	MQM	Yes	4	4,505	72,080
WMT MQM 2023	En→De, Zh→En, He→En	MQM	Yes	5	2,457	35,472
WMT MQM 2024	En→De, En→Es, Ja→Zh	MQM	Yes	4	1,667	24,461
MLQE-PE	En→De, Zh, Ja, Cs ǀ Ru, Ro, Et, Ne, Si, Ps, Km → En	DA + Post-editing	Yes	2	9,000 (per lang)	67,000
SCB-MT-EnTh-2020	En↔Th	-	Yes	6	1M	1M
MEET-MR (Ours)	En→Th	MQM + Ranking	Yes	9	2,142	20,100

Table 1: Comparison of our dataset with existing human-evaluated MT datasets. Note: WMT MQM sizes refer to the subsets annotated with MQM, not the full General Task test sets.

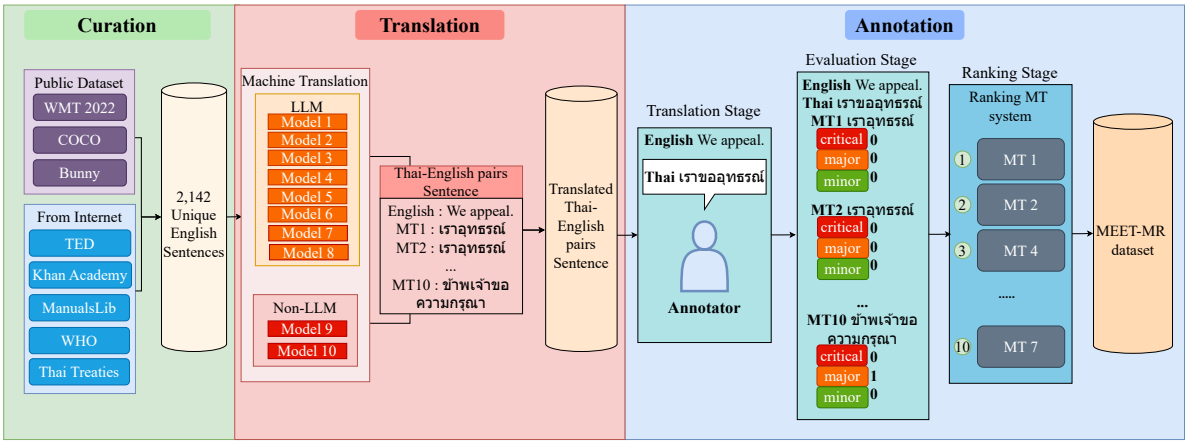

Figure 1: The dataset construction, including data curation, translation generation, and human annotation.

Domain	#Sentences
Education	296
E-commerce	292
Conversation	284
Social	273
News	246
Medical	215
Treaty	211
Smart City	207
Manual	118
Total	**2,142**

Table 2: Distribution of source sentences in MEET-MR across domains.

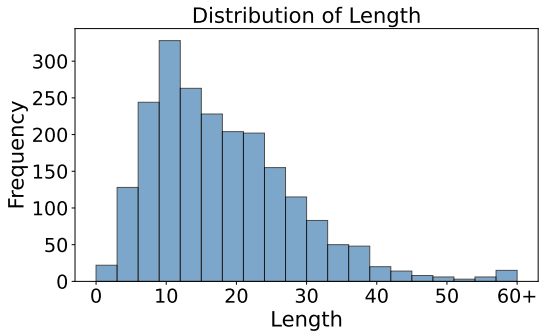

Figure 2: Distribution of segment lengths in MEET-MR.

2.4 Data Cleansing and Quality Validation

Despite using qualified annotators, human errors and inconsistencies are inevitable. To improve reliability, we applied two complementary validation methods.

Segment correlation consistency. We first convert MQM error counts into a scalar MQM score following Rei et al. (2020):

$$q = 1 - \frac{\text{minor} + 5 \cdot \text{major} + 10 \cdot \text{critical}}{\text{length}}. \quad (1)$$

For each source sentence with at most 10 MT outputs, we computed Kendall's τ between q and the annotator-provided rankings. Sentences with low Kendall's τ were flagged for manual review, since they indicated possible annotation inconsistencies.

Pairwise inconsistency detection. We also examined pairs of MT outputs for the same source. For translations i and j we defined:

$$A(i,j) = |(q_i - q_j) + (r_i - r_j)|, \quad (2)$$

Error Type	Mean	SD	Mean Rate (%)
Critical	0.303	2.082	2.07
Major	0.345	0.711	2.60
Minor	0.580	1.044	4.17

Table 3: Distribution statistics of MQM error types in the English–Thai dataset. Mean and SD are computed per sentence; Mean Rate represents the average percentage of errors relative to sentence length.

where q is MQM score and r is the rank. Intuitively, if a translation receives a higher score but a worse rank (or vice versa), $A(i, j)$ will be large. We flagged pairs with high $A(i, j)$ as potential contradictions for further inspection.

Flagged items were reviewed by a second annotator. Obvious errors were corrected and ambiguous cases were re-annotated. This procedure systematically reduced the annotation noise while preserving the majority of reliable annotations.

Annotator Agreement To test the validity of our idea of using a ranking scheme for scoring, we conducted a small experiment. We asked another set of annotators to re-annotate a small subset of the data selected from each domain. We then measured annotator consistency using average Kendall's τ. The MQM scores achieved an average τ of 0.462, while the ranking scheme achieved 0.502 (see Section 3.2). This is because, even with the same MQM guidelines, annotators may differ in how they judge error severity, which can greatly affect MQM scores. In contrast, ranking reflects overall human preference-annotators generally agree on which translations are better, so their relative ordering remains more stable. This observation is shown in Appendix D.1.

Comparison to Existing Datasets The comparison of between our dataset and other widely benchmark are shown in Table 1. While WMT MQM (Freitag et al., 2021a) and MLQE-PE (Fomicheva et al., 2020) provide valuable human evaluation and quality estimation resources, they do not include the English-Thai language pair. In contrast, SCB-MT (Lowphansirikul et al., 2021) offers large-scale English-Thai parallel data for MT training but does not incorporate human evaluation. Our dataset bridges this gap by providing English-Thai translations that combine domain diversity with fine-grained human evaluation based on both MQM and Relative Ranking frameworks.

3 Experimental Setup

In this section, we describe how to use our dataset to benchmark a Quality Estimation (QE) model, which predicts translation quality without relying on human reference translations.

3.1 Train-Test Split

The MEET-MR dataset was partitioned by English source sentences, ensuring that all translations remained within the same split. The data was divided into training, validation, and test sets using an 80:10:10 ratio. Each subset was stratified by domain, preserving the overall domain distribution across the splits.

3.2 Evaluation Metrics

We used Kendall's τ (Kendall, 1938) as an evaluation metric to measure the consistency between model predictions and human annotations. For each unique source segment, we calculated the correlation between model prediction and human annotation. This follows Freitag et al. (2021a,b) which has used Kendall's τ for sentence-level MQM evaluation. Kendall's τ is defined as follows:

$$\tau = \frac{C - D}{C + D}, \tag{3}$$

where C denotes the number of concordant pairs, for which the relative ordering of model predictions and human judgments agrees, and D denotes the number of discordant pairs, for which the orderings disagree.

We have two evaluation methods. The first is MQM τ correlation, where compared against annotator's MQM. The second is Rank τ, where compared against annotator's Relative Ranking. The model performance was represented by these two metrics, providing a balanced assessment of how well the model represents translation quality and translation preferences.

3.3 Models

In this study, we focus on quality estimation by employing three approaches for translation quality estimation: (1) COMET-21 (Rei et al., 2021), which was finetuned on the MEET-MR dataset; (2) COMET-kiwi (Rei et al., 2022), another COMET-based variant also finetuned on MEET-MR dataset; and (3) GEMBA-MQM (Kocmi and Federmann, 2023), a prompted LLM to assess translation quality in both zero-shot and few-shot settings. For the

Model	MQM $\tau \uparrow$	Rank $\tau \uparrow$
Pretrained		
COMET-21	0.272	0.290
COMET-kiwi	0.362	0.383
Finetuned		
COMET-21	0.323*	0.353*
COMET-kiwi	0.402*	0.415*
LLMs zero-shot		
MQM Claude	0.423	0.405
MQM Gemini	0.463	0.455
Ranking Claude	0.386	0.369
Ranking Gemini	0.474	0.485
LLMs few-shot		
MQM Claude	0.402	0.385
MQM Gemini	0.466	0.456
Ranking Claude	0.417*	0.412*
Ranking Gemini	0.483	0.500

Table 4: Kendall's τ of models with human average MQM scores and system ranking. An asterisk (*) indicates that finetuned or few-shot of that method is significantly better than pretrained or zero-shot using Fisher transformation ($p < 0.05$).

MQM few-shot setup, we randomly included 20 example sentences from each domain. We also prompted the LLMs to rank translations across each source segment according to overall quality. In this Ranking few-shot setup, we selected two unique source sentences from each domain, and each source sentence has at most 10 MT outputs. More prompt details in Appendix C.

Finetuning COMET on the MEET-MR dataset required approximately five hours for a single training run on a single NVIDIA A100 40GB GPU. Optimization was performed using the AdamW optimizer (Loshchilov and Hutter, 2019). Hyperparameters were tuned by finetuning the model for five epochs, using Kendall's τ as the optimization objective to select the best configuration.

4 Results and Discussion

Both finetuned COMET-21 and COMET-kiwi outperform the COMET pretrained, as shown in Table 4. The finetuned model shows stronger alignment with human MQM scores than the pretrained version, suggesting that finetuning helps the model to better capture Thai vocabulary and contextual nuances. For GEMBA-MQM, we compare Relative Ranking performance between zero-shot and few-shot settings. The zero-shot model tends to group translations with similar wording, showing limited sensitivity to subtle differences, whereas the few-shot model better reflects human preferences, producing rankings more consistent with human judgements.

In the zero-shot setting, GEMBA-MQM outperformed the COMET models. However, few-shot examples only improves in the Gemini case. Further analysis reveals that Claude did not make use of the provided examples. When we input the exact segments already included in the prompt, Claude answered incorrectly 21 out of 180 times. On the other hand, few shots improves the LLM in the Ranking setup for both LLMs.

This inconsistent behavior might be due to how ranking annotations can capture nuances better than MQM annotations. To demonstrate how our annotation scheme captures human preferences even when translation errors are minimal, we analyze an interesting example. Although only one output contains a minor error in MQM framework, the translations can still be ranked across 10 distinct levels of quality. This suggests that Relative Ranking or human preference judgements can capture nuances of quality beyond what the MQM framework explicitly measures. See Appendix D.2 for more details.

5 Conclusion

We present a new English-Thai dataset, which we use to finetune quality estimation models and benchmark against an LLM-based evaluator. Experimental results on our test set show that while LLM-based evaluators achieve the highest performance, they remain closed-source and require high computational resources, limiting their accessibility and reproducibility. In contrast, our finetuned COMET models significantly narrow the performance gap, offering a more lightweight and open alternative. With more extensive and diverse training data, these models have the potential to match the performance of LLM-based evaluators, facilitating efficient and open assessment of English-Thai quality estimation.

Limitations

Our dataset has several limitations. First, it is restricted to short text segments and does not account for discourse-level phenomena. Second, it only covers sentence-level evaluation without word-level annotations. Third, it focuses solely on English–Thai translation, limiting generalizability to other language pairs. Fourth, while annotated by qualified linguists, the quality of labels may vary due to inevitable subjectivity in human evaluation.

Future work can address these issues by collecting data from more diverse domains to enhance representativeness and by involving a larger pool of annotators to further reduce subjectivity and improve annotation consistency.

References

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106, Columbus, Ohio. Association for Computational Linguistics.

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.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021a. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021b. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Yvette Graham, Timothy Baldwin, and Nitika Mathur. 2015. Accurate evaluation of segment-level machine translation metrics. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1183–1191, Denver, Colorado. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Muyang He, Yexin Liu, Boya Wu, Jianhao Yuan, Yueze Wang, Tiejun Huang, and Bo Zhao. 2024. Efficient multimodal learning from data-centric perspective. *arXiv preprint arXiv:2402.11530*.

Maurice G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 Conference on Machine Translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Gemba-mqm: Detecting translation quality error spans with gpt-4. In *Proceedings of the Eighth Conference on Machine Translation*, Singapore. Association for Computational Linguistics.

Yeskendir Koishekenov, Alexandre Berard, and Vassilina Nikoulina. 2023. Memory-efficient NLLB-200: Language-specific expert pruning of a massively multilingual machine translation model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3567–3585, Toronto, Canada. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Arle Lommel, Aljoscha Burchardt, Maja Popović, Kim Harris, Eleftherios Avramidis, and Hans Uszkoreit. 2014. Using a new analytic measure for the annotation and analysis of MT errors on real data. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, pages 165–172, Dubrovnik, Croatia. European Association for Machine Translation.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Lalita Lowphansirikul, Charin Polpanumas, Attapol T. Rutherford, and Sarana Nutanong. 2021. A large english–thai parallel corpus from the web and machine-generated text. *Language Resources and Evaluation*, 56(2).

Raymond Ng, Thanh Ngan Nguyen, Yuli Huang, Ngee Chia Tai, Wai Yi Leong, Wei Qi Leong, Xianbin Yong, Jian Gang Ngui, Yosephine Susanto, Nicholas Cheng, Hamsawardhini Rengarajan, Peerat Limkonchotiwat, Adithya Venkatadri Hulagadri, Kok Wai Teng, Yeo Yeow Tong, Bryan Siow, Wei Yi Teo, Wayne Lau, Choon Meng Tan, and 12 others. 2025.

Sea-lion: Southeast asian languages in one network. *Preprint*, arXiv:2504.05747.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, USA. Association for Computational Linguistics.

Wannaphong Phatthiyaphaibun, Surapon Nonesung, Patomporn Payoungkhamdee, Peerat Limkonchotiwat, Can Udomcharoenchaikit, Jitkapat Sawatphol, Chompakorn Chaksangchaichot, Ekapol Chuangsuwanich, and Sarana Nutanong. 2024. Wangchan-Lion and WangchanX MRC Eval. *Preprint*, arXiv:2403.16127.

Kunat Pipatanakul, Phatrasek Jirabovonvisut, Potsawee Manakul, Sittipong Sripaisarnmongkol, Ruangsak Patomwong, Pathomporn Chokchainant, and Kasima Tharnpipitchai. 2023. Typhoon: Thai large language models. *arXiv preprint arXiv:2312.13951*.

Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. Are references really needed? unbabel-IST 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online. 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.

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.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Lucia Specia, Marina Fomicheva, Francisco Guzmán, Frédéric Blain, Erick Fonseca, Shuo Sun, Vishrav Chaudhary, Nina Lopatina, and André F. T. Martins. 2022. Mlqe-pe: A multilingual quality estimation and post-editing dataset. In *Proceedings of the 13th International Conference on Language Resources and Evaluation (LREC 2022)*.

Ye Kyaw Thu, Win Pa Pa, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. 2016. Introducing the Asian language treebank (ALT). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1574–1578, Portorož, Slovenia. European Language Resources Association (ELRA).

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*.

Sumeth Yuenyong, Kobkrit Viriyayudhakorn, Apivadee Piyatumrong, and Jillaphat Jaroenkantasima. 2025. Openthaigpt 1.5: A thai-centric open source large language model. *Preprint*, arXiv:2411.07238.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Data sources

Our dataset was collected across nine primary domains ensuring diversity across formal, informal, and technical contexts. We curated the original English sentences from publicly available resources to ensure compliance with data usage rights, and all content was manually screened for inappropriate material.

Open-source Datasets Many datasets contain English sentences, regardless of whether they were originally designed for machine translation tasks. We collected utterances from the training data of WMT MQM 2022 (Kocmi et al., 2022), covering domains such as *conversational text*, *social media*, *news articles*, and *e-commerce content*. In addition to purely textual datasets, we also incorporated image–text pair datasets to reflect the image captioning task. Specifically, we used data from the Bunny 1.0 dataset (He et al., 2024), drawn from both its pretraining and finetuning splits, and the COCO dataset (Lin et al., 2014), using the validation split from the 2017 COCO image captioning dataset.

Public websites To expand the variety of source materials, we focused on the domains of *education*, *technical manuals*, *smart city content*, and *medical terminology*. For the education domain, we extracted data from the TED[3] and Khan Academy[4] websites, which are well-known for self-directed learning and educational content. The sentences were collected from English subtitles of YouTube videos. For TED, we selected 125 sentences from both TED Talks and TED-Ed videos to ensure diversity in content and style, extracting at least one sentence from each clip. For Khan Academy, we selected 230 sentences from the video subtitles of education course clips, focusing on mathematics and science playlists. Science sentences were specifically chosen from physics, chemistry, and biology courses targeted at the high school level.

For the manuals, we collected documents from the Manualslib website[5], covering domains such as furniture, vehicles, and electronic devices. For the official texts, we used treaties obtained from the Thai official treaty website[6]. For the medical domain, we used texts from the World Health Organization (WHO) health topics website[7], focusing on country-specific issues related to diseases and healthcare, ensuring the inclusion of medical terminology.

The MEET-MR dataset will be publicly released under the Creative Commons Attribution 4.0 International (CC-BY-4.0) license.

B Annotation User Interface and Annotation Guideline

The data labeling process consisted of three main stages: translation, evaluation, and ranking. In the translation stage (Figure 3), the system displayed the source text in English, and annotators were instructed to produce high-quality translations in Thai. In the evaluation stage (Figure 4), annotators were asked to assess the outputs from different machine translation systems by identifying and categorizing translation errors into three severity levels: minor, major, and critical which are defined by

- **Minor:** errors as those that do not significantly alter meaning (e.g., minor punctuation issues)

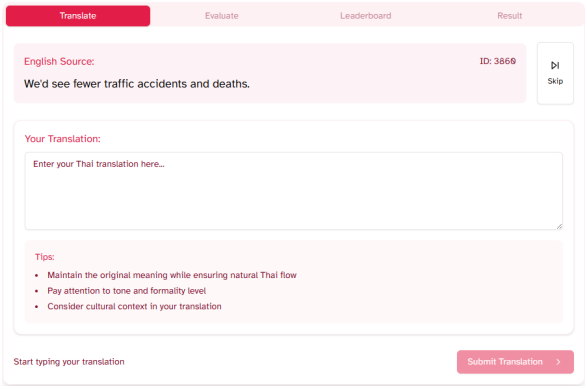

Figure 3: Translation Stage

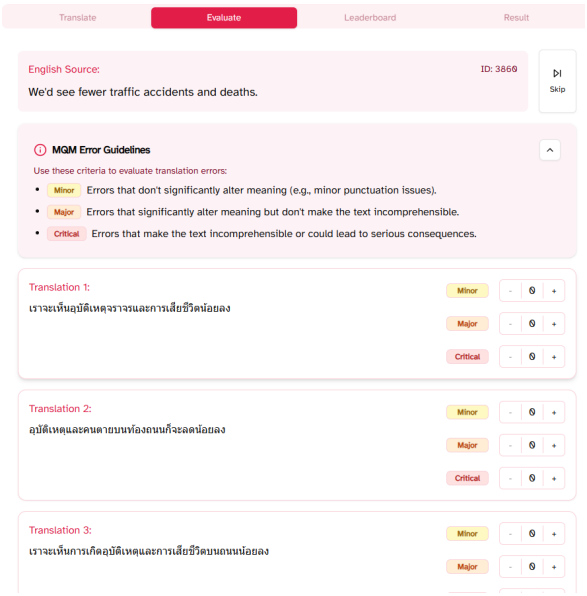

Figure 4: Evaluation Stage

- **Major:** errors as those that significantly alter meaning without rendering the text incomprehensible

- **Critical:** errors as those that make the text incomprehensible or could lead to serious consequences.

In the ranking stage (Figure 5), annotators were instructed to rank the outputs of the machine translation systems according to their overall quality, with the option to assign tied rankings when systems were judged to be of equivalent quality.

C LLM Prompt

LLM-based evaluators were performed using **Gemini-Pro-2.5**[8] and **Claude-Sonnet-4.0**[9].

[3] www.ted.com/talks, www.youtube.com/@TEDEd
[4] www.youtube.com/@khanacademy
[5] www.manualslib.com
[6] https://treaties.mfa.go.th
[7] www.who.int/thailand/health-topics

[8] Version: Gemini-2.5-Pro-Stable-06-17
[9] Version: Claude-Sonnet-4-20250514

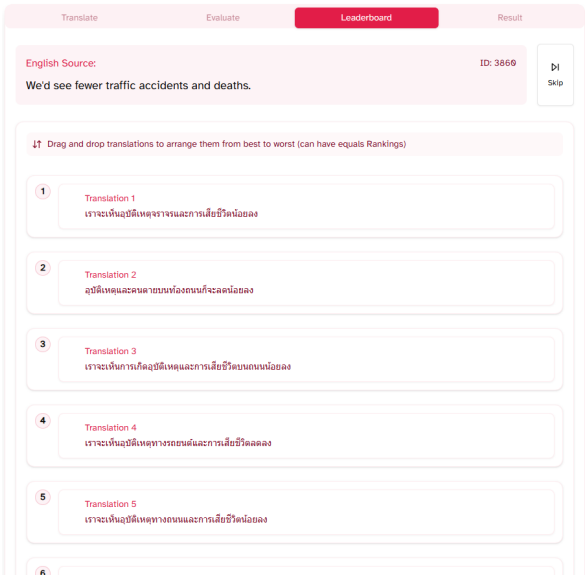

Figure 5: Ranking Stage

C.1 MQM-Style Error Prediction

Figures 6 – 7 present the prompt templates used for the inference of **Gemini-Pro-2.5** and **Claude-Sonnet-4.0** in both zero-shot and few-shot settings, correspondingly. An MQM framework was implemented for quality estimation of translations. The prompt instructed the model to perform quality estimation for machine translation outputs, systematically identify translation errors, categorize them according to predefined taxonomies, and report the results as a triplet `[count_minor, count_major, count_critical]`.

C.2 Translation Ranking

Figures 8 – 9 present the prompt templates used for the inference of **Gemini-Pro-2.5** and **Claude-Sonnet-4.0** in both zero-shot and few-shot settings, correspondingly. A comparative ranking prompt was designed to rank machine translation outputs for a given English source sentence. The model was required to return a JSON object assigning integer ranks, with 1 indicating the highest quality and 10 the lowest.

D Additional samples

D.1 Re-annotation Sample

Table 5 presents a re-annotated example comparing the annotations from two independent annotators. Despite following the same MQM guidelines, the annotators differ in how they judge error severity across machine translation outputs, which leads to discrepancies in MQM scores. However, relative ranking reflects human preference, making it less sensitive to such disagreements. In this sample, MQM scores achieved Kendall's τ of 0.235, while the relative ranking achieved Kendall's τ of 0.816.

D.2 Translation sample

Table 6 presents a machine translation outputs evaluated by human in MQM framework and relative ranking. The source sentence is:

> *The total number of vaccines injected in the country reached 109,990,742 doses.*

D.3 Model sample

Table 7 presents a comparison between COMET-21 and human evaluations. The results show that the COMET-21 pretrained model, which is trained on high-resource language pairs, struggles to capture complex Thai vocabulary and contextual nuances, leading to consistently lower prediction scores across all translation sentences. In contrast, the COMET-21 finetuned model shows stronger alignment with human MQM scores.

Table 8 presents a comparison between COMET-kiwi and human evaluations. The COMET-Kiwi pretrained model captures basic Thai vocabulary but fails to capture the naturalness of Thai sentences, assigning high scores to sentences that are unnatural in Thai. In contrast, the COMET-Kiwi finetuned model aligns with human annotations and better captures sentence context.

Table 9 presents a comparison of zero-shot and few-shot settings in the Claude ranking model, along with the corresponding human rankings.

MQM Prompt Template

You are an **English-Thai translation evaluator**. Compare an English sentence and its Thai translation. **Count translation errors** based on natural Thai usage and meaning.

Based on the source segment and machine translation, identify error types in the translation and classify them. **The categories of errors are: accuracy (addition, mistranslation, omission, untranslated text), fluency (character encoding, grammar, inconsistency, punctuation, register, spelling), locale convention (currency, date, name, telephone, or time format), style (awkward), terminology (inappropriate for context, inconsistent use), non-translation, other, or no-error**.

Each error is classified as one of three categories: critical, major, and minor. Critical errors inhibit comprehension of the text. Major errors disrupt the flow, but what the text is trying to say is still understandable. Minor errors are technically errors, but do not disrupt the flow or hinder comprehension.

Output format (exact): `[count_minor, count_major, count_critical]`

Rules: Prioritize **natural Thai expressions and colloquial usage**.

Consider **cultural context and local idioms** where appropriate.

Focus on **natural, colloquial Thai**. **Do not include formal polite endings**.

Only output the counts. Do not explain.

Source Text (EN): {en}
Translated Text (TH): {th}

Figure 6: MQM-style evaluation prompt template.

Source	The interim prime minister, Abdul Hamid Dbeibah and warlord Khalifa Haftar, the head of the self-styled Libyan National Army, have meanwhile been approved by the commission but subsequently appealed by other parties.							
Reference	นายอับดุล ฮามิด ดีเบย์บาห์ ผู้รักษาราชการแทนนายกรัฐมนตรี และขุนศึก คาลิฟา ฮาฟทาร์ ผู้นำกองกำลังแห่งชาติลิเบียที่ก่อตั้งขึ้นเอง ถูกยื่นอุทธรณ์จากพรรคอื่น ๆ หลังจากที่อนุมัติจากคณะกรรมการ							
Machine Translation		ranking_1	critical_1	major_1	minor_1	ranking_2	critical_2	major_2	minor_2
นายกรัฐมนตรีชั่วคราว อับดุล ฮามิด ไบบาห์ และขุนศึก คาลิฟา ฮาฟตาร์ หัวหน้ากองทัพแห่งชาติลิเบีย ได้รับการอนุมัติจากคณะกรรมาธิการ แต่ต่อมาถูกยื่นอุทธรณ์โดยฝ่ายอื่นๆ		1	0	0	1	1	0	2	0
นายกรัฐมนตรีชั่วคราว อับดุล ฮามิด ด์ไบบะฮ์ และนายพลฝ่ายทหาร ขาลิฟะ ฮัฟตาร์ ผู้นำกองทัพแห่งชาติลิเบียที่ประกาศตนเอง ได้รับการอนุมัติจากคณะกรรมการ แต่ต่อมากิถูกอุทธรณ์โดยฝ่ายอื่นๆ		5	0	0	2	1	0	2	0
นายกรัฐมนตรีชั่วคราว อาบดุล ฮามิด ด์บัยบาห์ และ/String ซึ่งเป็นผู้นำกองทัพลิเบียแห่งชาติที่เรียกตนเองว่าuite掰岡_fid โดยทั้งคู่ได้รับการอนุมัติจากรัฐสภาระหว่างนี้แต่ถูกยื่นอุทธรณ์โดยฝ่ายอื่นๆ ที่ไม่ yaptığı pleased Note: There's a bit of a mix-up with the string at the mention of Khalifa Haftar, which seems to have been translated imperfectly. Here's the corrected version: นายกรัฐมนตรีชั่วคราว อาบดุล ฮามิด ด์บัยบาห์ และเคห์ลีฟา ฮะฟตาร์ ผู้นำกองทัพลิเบียแห่งชาติที่เรียกตนเองว่า ได้รับการอนุมัติจากรัฐสภาระหว่างนี้แต่ถูกยื่นอุทธรณ์โดยฝ่ายอื่นๆ		9	26	1	1	10	10	0	0
นายกรัฐมนตรีชั่วคราว อับดุล ฮามิด ดาเบิบาห์ และนายพลคาลิฟา ฮาฟตาร์ ผู้นำกองทัพแห่งชาติลิเบียที่ประกาศตนเอง ได้รับการอนุมัติจากคณะกรรมการ แต่ภายหลังมีการอุทธรณ์จากฝ่ายอื่น ๆ		3	0	0	0	1	0	0	1
นายอับดุลฮามิด ดเบียบาห์ นายกรัฐมนตรีรักษาการ และนายคาลิฟา ฮาฟตาร์ หัวหน้ากองทัพแห่งชาติลิเบีย (ซึ่งจัดตั้งขึ้นเอง) ได้รับการอนุมัติจากคณะกรรมการแล้ว แต่ต่อมาถูกอุทธรณ์โดยฝ่ายอื่นๆ \n\n		0	0	0	2	1	0	0	1
นายกรัฐมนตรีรักษาการ อาบดุล ฮามิด ด์บิบา และขุนศึกคาราอีฟา ฮัฟตาร์ หัวหน้ากองทัพแห่งชาติลิเบียที่เรียกตัวเอง ได้รับการอนุมัติจากคณะกรรมาธิการ แต่ต่อมากลุ่มฝ่ายอื่นได้ยื่นอุทธรณ์		4	1	1	2	1	0	0	2
นายกรัฐมนตรีรักษาการ อับดุล ฮามิด ดไบบาห์ และผู้นำกองกำลังติดอาวุธ คาลิฟา ฮัฟตาร์ ซึ่งเป็นผู้นำกองทัพแห่งชาติลิเบียที่ตั้งขึ้นเอง ได้รับการรับรองจากคณะกรรมการแล้ว แต่ภายหลังถูกคัดค้านโดยฝ่ายอื่นๆ		2	0	1	2	1	0	0	1
ในขณะเดียวกัน คณะกรรมการได้อนุมัตินายกรัฐมนตรีชั่วคราว อับดุล ฮามิด ดบิบาห์ และนายกสงคราม คาลิฟา ฮาฟตาร์ ผู้ใหญ่ของกองทัพแห่งชาติลิเบียที่เรียกตัวเองว่าเป็นกองทัพแห่งประเทศลิเบีย แต่ภายหลังก็ได้รับการร้องเรียนจากพรรคอื่น ๆ		8	0	3	1	9	0	1	1
นายอับดุลฮามิด ดีบาติ ผู้นำรัฐบาลเฉพาะกาลและนายคาลิฟา ฮาฟตาร์ หัวหน้ากองกำลังติดอาวุธลิเบีย ได้รับการรับรองจากคณะกรรมการแล้ว แต่ในภายหลังได้ยื่นอุทธรณ์กับหน่วยงานอื่น ๆ		7	0	1	1	8	1	3	0
นายกรัฐมนตรีรักษาการ อับดุล ฮามิด ดิบายาห์ และผู้นำกลุ่มกบฏ คาลิฟา ฮาฟตาร์ หัวหน้าของกองทัพชาติอิสระลิเบีย ได้รับการอนุมัติจากคณะกรรมการแต่ถูกอุทธรณ์โดยฝ่ายอื่น		6	0	2	0	7	0	0	3

Table 5: Example re-annotation of a single source segment, comparing error severity and ranking annotations between Annotator 1 and Annotator 2.

MQM Prompt Template

You are an **English-Thai translation evaluator**. Compare an English sentence and its Thai translation. **Count translation errors** based on natural Thai usage and meaning.

Based on the source segment and machine translation, identify error types in the translation and classify them. **The categories of errors are: accuracy (addition, mistranslation, omission, untranslated text), fluency (character encoding, grammar, inconsistency, punctuation, register, spelling), locale convention (currency, date, name, telephone, or time format), style (awkward), terminology (inappropriate for context, inconsistent use), non-translation, other, or no-error**.

Each error is classified as one of three categories: critical, major, and minor. Critical errors inhibit comprehension of the text. Major errors disrupt the flow, but what the text is trying to say is still understandable. Minor errors are technically errors, but do not disrupt the flow or hinder comprehension.

Output format (exact): `[count_minor, count_major, count_critical]`

Examples:

EN: {en_1}
TH: {th_1}
OUTPUT: `[minor, major, critical]`

. . . (2*10 sentences from domains)

EN: {en_n}
TH: {th_n}
OUTPUT: `[minor, major, critical]`

Rules: Prioritize **natural Thai expressions and colloquial usage**.

Consider **cultural context and local idioms** where appropriate.

Focus on **natural, colloquial Thai**. **Do not include formal polite endings**.

Only output the counts. Do not explain.

Source Text (EN): {en}
Translated Text (TH): {th}

Figure 7: MQM-style few-shot evaluation prompt template.

Translation Ranking Prompt Template

You are an **English-Thai translation evaluator**. Compare an English sentence and 10 Thai translations. **Rank each translation** based on naturalness, accuracy, and cultural appropriateness.

Ranking Objective: Determine which translations most faithfully convey the meaning of the English source while sounding natural, fluent, and appropriate in Thai.

Ranking Scale:
– 1: Perfect or near-perfect translation. Flawless in meaning, natural, and fluent.
– 10: Completely incorrect, nonsensical, or fails to convey the original meaning.

Evaluation Criteria (in order of importance):
Accuracy: Does the translation convey the full and precise meaning of the English source? Are there any mistranslations, omissions, or additions?
Fluency and Naturalness: Does it sound like natural, modern, colloquial Thai? Is the sentence smooth, idiomatic, and grammatically correct? Avoid overly literal or awkward phrasing.
Cultural Appropriateness: Does it handle idioms, expressions, or context correctly for a Thai audience?

Rules:
Prioritize **natural, colloquial Thai expressions**. **Ignore polite particles** and do not penalize for them. **Ties are allowed**: assign the same rank to translations of indistinguishable quality. **Only output the ranks as a JSON object**. Do not provide explanations, comments, or markdown formatting.

Output Format (Exact):
`{id1: rank, id2: rank, ..., id10: rank}`
Source Text (EN): {en}
Translated Texts (TH): {List of 10 sentences, each have unique `id` and `text`}.

Figure 8: Thai translation ranking style prompt template.

Translation Ranking Prompt Template

You are an **English-Thai translation evaluator**. Compare an English sentence and 10 Thai translations. **Rank each translation** based on naturalness, accuracy, and cultural appropriateness.

Ranking Objective: Determine which translations most faithfully convey the meaning of the English source while sounding natural, fluent, and appropriate in Thai.

Ranking Scale:
– 1: Perfect or near-perfect translation. Flawless in meaning, natural, and fluent.
– 10: Completely incorrect, nonsensical, or fails to convey the original meaning.

Evaluation Criteria (in order of importance):
Accuracy: Does the translation convey the full and precise meaning of the English source? Are there any mistranslations, omissions, or additions?
Fluency and Naturalness: Does it sound like natural, modern, colloquial Thai? Is the sentence smooth, idiomatic, and grammatically correct? Avoid overly literal or awkward phrasing.
Cultural Appropriateness: Does it handle idioms, expressions, or context correctly for a Thai audience?

Examples:
EN: {en_1}
TH candidates: {id1: th_1, ..., id10: th_10}
OUTPUT: `{id1: rank, ..., id10: rank}`
. . . (2*10 sentences from domains)
EN: {en_n}
TH candidates: {id1: th_1, ..., id10: th_10}
OUTPUT: `{id1: rank, ..., id10: rank}`

Rules:
Prioritize **natural, colloquial Thai expressions**. **Ignore polite particles** and do not penalize for them. **Ties are allowed**: assign the same rank to translations of indistinguishable quality. **Only output the ranks as a JSON object**. Do not provide explanations, comments, or markdown formatting.

Output Format (Exact):
`{id1: rank, id2: rank, ..., id10: rank}`
Source Text (EN): {en}
Translated Texts (TH): {List of 10 sentences, each have unique `id` and `text`}.

Figure 9: Thai translation ranking style few-shot prompt template.

Source	The total number of vaccines injected in the country reached 109,990,742 doses.			
Reference	จำนวนผู้ฉีดวัคซีนในประเทศรวมทั้งสิ้นถึง 109,990,742 โดสแล้ว			
Machine Translation	ranking	critical	major	minor
จำนวนวัคซีนที่ฉีดให้ประชาชนทั่วประเทศรวมทั้งสิ้น 109,990,742 โดส	0	0	0	0
จำนวนวัคซีนที่ฉีดในประเทศแล้วรวมทั้งสิ้น 109,990,742 โดส	1	0	0	0
จำนวนรวมของวัคซีนที่ถูกฉีดในประเทศถึง 109,990,742 โดส	2	0	0	0
ยอดฉีดวัคซีนในประเทศทะลุ 109,990,742 โดส	3	0	0	0
จำนวนวัคซีนที่ฉีดในประเทศรวมกันถึง 109,990,742 โดสแล้ว	4	0	0	0
จำนวนการฉีดวัคซีนในประเทศรวมถึง 109,990,742 โดส	5	0	0	0
จำนวนวัคซีนที่ฉีดในประเทศถึง 109,990,742 โดสแล้วครับ/ค่ะ	6	0	0	1
จำนวนวัคซีนที่ฉีดในประเทศทั้งหมดได้ถึง 109,990,742 โดสแล้ว	7	0	0	0
จำนวนวัคซีนที่ฉีดในประเทศทั้งหมดเพิ่มขึ้นเป็น 109,990,742 โดส	8	0	0	0
จำนวนวัคซีนที่ฉีดในประเทศทั้งหมดอยู่ที่ 109,990,742 โดส	9	0	0	0

Table 6: Example of relative ranking annotations for a single source segment. Red text marks a minor MQM error, and orange text highlights a word failing to convey the intended sense of “reach”.

Source Sentence	1. The premises of the mission shall be inviolable. The agents of the receiving State may not enter them, except with the consent of the head of the mission.		
Reference Sentence	1. อาคารและสถานที่ของคณะผู้แทนจะถูกละเมิดมิได้ ตัวแทนของรัฐผู้รับจะไม่สามารถเข้าไปยังสถานที่นั้นได้ เว้นแต่ได้รับความยินยอมจากหัวหน้าคณะผู้แทน		
Machine Translation	Human score	COMET-21 pretrained score	COMET-21 finetuned score
1. สถานที่ปฏิบัติภารกิจจะถูกละเมิดมิได้ ตัวแทนของรัฐผู้รับจะเข้าไปไม่ได้ เว้นแต่จะได้รับความยินยอมจากหัวหน้าคณะผู้แทน	0.793	0.121	0.830
1. สถานที่ของพันธกิจต้องได้รับการคุ้มครองไม่ให้ถูกล่วงละเมิด เจ้าหน้าที่ของรัฐที่รับรองไม่สามารถเข้าไปในสถานที่ดังกล่าวได้ เว้นแต่จะได้รับความยินยอมจากหัวหน้าพันธกิจ	0.655	0.130	0.844
1	0.310	0.082	0.251
สถานที่ตั้งของคณะผู้แทนจะต้องไม่ถูกล่วงละเมิด เจ้าหน้าที่ของรัฐผู้รับไม่สามารถเข้าไปในสถานที่ดังกล่าวได้ เว้นแต่จะได้รับความยินยอมจากหัวหน้าคณะผู้แทน	0.966	0.124	0.851
1. สถานทูตต้องเป็นที่เคารพและไม่อาจเข้าถึงได้ เว้นแต่จะได้รับความยินยอมจากหัวหน้าคณะผู้แทนทางทูต	0.621	0.132	0.770
1. สถานที่ของพันธกิจจะต้องไม่ถูกละเมิด เจ้าหน้าที่ของรัฐที่รับพันธกิจไม่สามารถเข้าไปในสถานที่ดังกล่าวได้ เว้นแต่จะได้รับความยินยอมจากหัวหน้าพันธกิจ	0.448	0.128	0.717
1. สถานที่ของภารกิจจะต้องไม่ถูกล่วงละเมิด เจ้าหน้าที่ของรัฐผู้รับไม่สามารถเข้าไปในสถานที่เหล่านั้นได้ เว้นแต่จะได้รับความยินยอมจากหัวหน้าภารกิจ	0.655	0.126	0.791
1. สถานที่ของภารกิจจะต้องไม่ถูกละเมิด เจ้าหน้าที่ของรัฐผู้รับจะไม่สามารถเข้ามาในสถานที่นั้นได้ ยกเว้นมีความยินยอมจากหัวหน้าภารกิจ	0.621	0.122	0.712
สถานที่ตั้งของสถานเอกอัครราชทูต/สถานทูตนั้นต้องเป็นที่ศักดิ์สิทธิ์และไม่อนุญาตให้เจ้าหน้าที่ของรัฐผู้รับเข้าไป เว้นแต่ได้รับความยินยอมจากหัวหน้าสถานเอกอัครราชทูต/สถานทูตนั้น	0.414	0.134	0.776
MQM score correlation		-0.2041	0.5443

Table 7: Example comparison of a single source segment across human evaluation, COMET-21 pretrained, and COMET-21 finetuned models. Note that "1" refers to the output generated by NLLB-200-1.3B (Koishekenov et al., 2023).

Source Sentence	A hydrating day & night cream that leaves thirsty skin soft, supple and refreshed.
Reference Sentence	ครีมให้ความชุ่มชื้นสำหรับกลางวันและกลางคืนที่ช่วยให้ผิวแห้งกลับมานุ่ม ยืดหยุน และรู้สึกสดชื่น

Machine Translation	Human score	COMET-kiwi pretrained score	COMET-kiwi finetuned score
ครีมน้ำในวันและคืน ที่ทำให้ผิวที่แห้งอ่อนนุ่ม และยืดหยุ่น	0.000	0.754	0.196
ครีมบำรุงผิวสูตรเพิ่มความชุ่มชื้น ใช้ได้ทั้งกลางวันและกลางคืน ช่วยให้ผิวที่แห้งกระหายกลับมานุ่มชุ่มชื้น เปล่งปลั่งสดใส	0.571	0.856	0.715
ครีมบำรุงผิวกลางวันและกลางคืนที่ช่วยให้ผิวแห้งกร้านนุ่ม ชุ่มชื้น และสดชื่น	1.000	0.824	0.795
ครีมบำรุงผิวหน้าและกลางคืนที่ให้ความชุ่มชื้น ทำให้ผิวที่ขาดน้ำรู้สึกนุ่มนวล ยืดหยุ่น และสดชื่น	0.286	0.849	0.596
ครีมบำรุงผิวกลางวันและกลางคืน ที่ช่วยให้ผิวแห้งนุ่ม น่าสัมผัส และรู้สึกสดชื่น	0.571	0.571	0.556
ครีมบำรุงผิวหน้าทั้งกลางวันและกลางคืนที่ทำให้ผิวที่แห้งแล้งกลับมาเนียนนุ่ม และสดชื่น	0.929	0.847	0.870
ครีมให้ความชุ่มชื้นสำหรับกลางวันและกลางคืนที่ช่วยให้ผิวที่กระหายน้ำนุ่ม อ่อนนุ่ม และสดชื่น	0.214	0.849	0.404
ครีมชโลมผิวสำหรับกลางวันและกลางคืนที่ช่วยเติมความชุ่มชื้นให้ผิวที่แห้งกระด้าง ทำให้ผิวนุ่ม ชุ่มชื่น และสดชื่นทันที	0.286	0.830	0.525
ครีมให้ความชุ่มชื้นสำหรับกลางวันและกลางคืนที่ทำให้ผิวที่ขาดน้ำนุ่ม ยืดหยุ่น และสดชื่น	1.000	0.855	0.851
ครีมบำรุงกลางวันและกลางคืน ที่ให้ความชุ่มชื้นแก่ผิวแห้งกร้าน ให้ผิวนุ่มนวล ชุ่มชื่น และรู้สึกสดชื่น	1.000	0.841	0.787
MQM score correlation		0.0943	0.7542

Table 8: Example comparison of a single source segment across human evaluation, COMET-kiwi pretrained, and COMET-kiwi finetuned models. Red text shows incorrect natural words used, and Orange text shows missing translations from the English source sentence.

Source	(vii) The term "executive head" means the principal executive official of the specialized agency in question
Reference	(๗) คำว่า “หัวหน้าบริหาร” หมายความถึง พนักงานฝ่ายบริหารชั้นหัวหน้าแห่งทบวงการชำนัญพิเศษที่อยู่ในปัญหานั้น

Machine Translation	Human ranking	Zero-shot ranking	Few-shot ranking
(vii) คำว่า "หัวหน้าฝ่ายบริหาร" หมายถึงเจ้าหน้าที่บริหารระดับสูงของหน่วยงานเฉพาะด้านที่เกี่ยวข้อง	3	3	6
และ ธ. การใช้คำว่า "หัวหน้าผู้บริหาร" หมายถึงเจ้าหน้าที่ผู้บริหารหลักของหน่วยงานเฉพาะทางที่เกี่ยวข้อง	7	7	9
คำว่า "หัวหน้าฝ่ายบริหาร" หมายถึง เจ้าหน้าที่บริหารระดับสูงสุดของทบวงการชำนัญพิเศษที่เกี่ยวข้อง	2	1	3
(vii) คำว่า "หัวหน้าฝ่ายบริหาร" หมายถึงเจ้าหน้าที่บริหารหลักของหน่วยงานเฉพาะทางที่เกี่ยวข้อง	3	2	2
(vii) คำว่า "หัวหน้าฝ่ายบริหาร" หมายถึงเจ้าหน้าที่บริหารระดับสูงขององค์การเฉพาะทางที่เกี่ยวข้อง	3	4	5
(vii) "หัวหน้าผู้บริหาร" หมายถึง เจ้าหน้าที่บริหารสูงสุดของหน่วยงานเฉพาะกิจนั้นๆ	6	5	7
(vii) คำว่า "หัวหน้าฝ่ายบริหาร" หมายความว่า เจ้าหน้าที่ฝ่ายบริหารสูงสุดของหน่วยงานพิเศษที่เกี่ยวข้อง	5	6	4
(vii) คำว่า "หัวหน้าฝ่ายบริหาร" หมายถึง เจ้าหน้าที่บริหารสูงสุดขององค์กรเฉพาะที่เกี่ยวข้อง	4	8	8
(vii) คำว่า "หัวหน้าผู้บริหาร" หมายถึง ผู้บริหารหลักของทบวงการชำนัญพิเศษดังกล่าว	1	9	1
Ranking Correlation		0.261	0.667

Table 9: Example comparison of a single source segment across human evaluation and GEMBA-MQM (Claude) rankings. The zero-shot setting assigns the lowest rank due to the unique Red texts which are absent in other translations, whereas the few-shot setting demonstrates improved overall ranking.