A Framework for Large-Scale Parallel Corpus Evaluation: Ensemble Quality Estimation Models Versus Human Assessment

Dmytro Chaplynskyi
Ukrainian Catholic University
lang-uk initiative
chaplynskyi.dmytro@ucu.edu.ua

Kyrylo Zakharov
UNHCR
kirillzakharov13@gmail.com

Abstract

We developed a methodology and a framework for automatically evaluating and filtering large-scale parallel corpora for neural machine translation (NMT). We applied six modern Quality Estimation (QE) models to score 55 million English-Ukrainian sentence pairs and conducted human evaluation on a stratified sample of 9,755 pairs. Using the obtained data, we ran a thorough statistical analysis to assess the performance of selected QE models and build linear, quadratic and beta regression models on the ensemble to estimate human quality judgments from automatic metrics. Our best ensemble model explained approximately 60% of the variance in expert ratings. We also found a non-linear relationship between automatic metrics and human quality perception, indicating that automatic metrics can be used to predict the human score. Our findings will facilitate further research in parallel corpus filtering and quality estimation and ultimately contribute to higher-quality NMT systems. We are releasing our framework, the evaluated corpus with quality scores, and the human evaluation dataset to support further research in this area.

1 Introduction

According to the Scaling Law (Kaplan et al., 2020), three basic ingredients are required to build a successful Large Language Model: the model's size, the amount of compute spent on training, and the size of the dataset. In this paper, we will focus on the latter. Indeed, the amount of text available is limited, and the limitation is even more visible for low-to-mid resource languages (see Zhong et al., 2024, Hasan et al., 2024). One way to tackle that problem is to translate a decent amount of text using Neural Machine Translation models, trading compute spent on inference to the data. Recent advances in the NMT models, such as NLLB (Team et al., 2022) and MadLad (Kudugunta et al., 2023), offer multilingual translation capabilities for hundreds of languages, building bridges to the low-resource languages.

Unfortunately, the measured quality of translation from English for these target languages is lower[1] than that for the popular pairs, such as English to German. This gap can be explained by the lack of training data (now for the NMT task) and the quality of the metrics. While metrics such as chrF (Popović, 2015) and BLEU (Papineni et al., 2002) are mechanistic and might not work well for fusional languages (Ma et al., 2019), others like Comet (Rei et al., 2020) or MetricX (Juraska et al., 2023) might not have enough knowledge about low-resourced languages, again, because of the underrepresentation.

If we look closer into the training of the NMT model, we might find the apparent abundance of Sent2Sent parallel corpora available online (Tiedemann, 2016). For example, when we began our research, the English to Ukrainian corpora had 97 million pairs, which now has around 158 million pairs[2].

However, a closer manual inspection reveals that at least part of the data is duplicated, garbled, or even obscene. Most importantly, one cannot assess the quality of the whole corpus at the scale needed to build an NMT model. These issues might visibly affect the quality of the models trained on this data (Sánchez-Cartagena et al., 2018).

As such, we identified the following research questions:

1. Can we automatically evaluate a big parallel corpora using State-of-the-Art quality estimation models?
2. How good are those models when compared to human evaluation?
3. Can we create an ensemble model to improve the quality of the evaluation?

[1] https://opus.nlpl.eu/dashboard/

[2] https://opus.nlpl.eu/results/en&uk/corpus-result-table

Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025), pages 73–85
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

To address these research questions, we created a methodology and a framework to collect parallel corpora at scale, deduplicate them, and score the individual sentence pairs using an ensemble of six Quality Estimation (QE) models that work in a multilingual setup. Additionally, we ran a human annotation of the stratified random sample, scoring 9775 pairs with the help of students of the linguistics faculty who are proficient in English and Ukrainian.

Using the obtained data, we ran a thorough statistical analysis to assess the performance of selected QE models and build linear, quadratic, and beta regression models on the ensemble to predict the human score.

Today, we are releasing the framework[3], the evaluated and deduplicated dataset of 55 million sentence pairs[4], and the data collected during the human evaluation. All the code, data, and instructions are published under permissive licenses to allow other scholars to reproduce the same workflow for other languages.

2 Related Work

The problem of filtering noisy parallel corpora has been addressed through several approaches: hybrid translation model-based filtering, machine learning classification, which frames filtering as a supervised task, multi-criteria heuristics combining statistical and neural techniques, and neural quality estimation models designed specifically for translation quality assessment.

2.1 Hybrid Translation Model-Based Filtering

Junczys-Dowmunt, 2018 proposed using dual conditional cross-entropy filtering, utilizing two inverse translation models trained on clean data to score each sentence pair. That work was limited to the English-Deutsch language pair.

2.2 Machine Learning Classification Approaches

Bicleaner (Sánchez-Cartagena et al., 2018) is another framework that discards sentences with visible flaws using handcrafted rules. It then applies classical ML algorithms and lexical similarity features to learn a score. Initially released for English-Deutsch, it now offers models for 33 language pairs[5].

Its experimental extension, bicleaner-ai (Zaragoza-Bernabeu et al., 2022), employs a transformer-based classifier and offers a smaller number of individual models for language pairs. It also offers a multilingual model that could potentially work with any language paired with English.

2.3 Multi-Criteria Heuristic Approaches

In our previous work (Paniv et al., 2024), we used a set of metrics, including the perplexity of both sentences and their similarity, calculated with the help of sentence transformers coupled with some hand-crafted rules to prepare the noisy corpus for training. In the final fine-tuning stage, we also utilized k-fold validation to filter a smaller dataset.

2.4 Neural Quality Estimation Models

Our current research operates three families of QE models from Unbabel and Google Research teams.

1. **COMET Family** (wmt22-cometkiwi-da by Rei et al., 2022, wmt23-cometkiwi-da by Rei et al., 2023) that combines COMET's architecture with the predictor–estimator setup of OpenKiwi, adding word-level tags and explanations achieving SOTA performance on Quality Estimation Shared Task. wmt23-cometkiwi-da models are built on a bigger backbone model and are available in different sizes.
2. **xCOMET** (Guerreiro et al., 2024), which integrates both sentence-level evaluation and error span detection capabilities and allows for a reference-free mode.
3. **MetricX Family** (MetricX-23 by Juraska et al., 2023 and MetricX-24 by Juraska et al., 2024), trained with a two-stage fine-tuning strategy on large human-labeled datasets. These models can also work in a reference-free mode.

While these approaches have shown promising results, most models have focused on high-resource language pairs or relied on clean parallel data for the training. Furthermore, comparisons between automatic quality estimation and human evaluation remain limited for the language pair of our interest. Our work addresses these gaps by evaluating multiple QE models against human judgments specifically for English-Ukrainian translation, providing

[3] https://github.com/lang-uk/vakula

[4] https://huggingface.co/datasets/lang-uk/FiftyFiveShades

[5] https://github.com/bitextor/bicleaner

insights into their performance for languages we need.

3 Methodology

To evaluate the quality of the English-Ukrainian parallel corpus at scale, we are proposing a pipeline which consists of the following stages:

1. Corpus collection
2. Automatic Quality Estimation with six QE models
3. Stratified sampling for the human evaluation
4. A solution for crowdsourced human evaluation
5. Statistical analysis of the results
6. Ensemble models fitting
7. Rescoring of the evaluated corpus using ensemble models

3.1 Corpus Collection

We used the already mentioned collection of parallel corpora from OPUS Open Parallel Corpora. It includes a handful of corpora for our interest's language pair and allows us to download them separately in the unified TMX[6] format. At the beginning of the research, it offered 97,062,370 pairs of sentences from 35 sources (see table 1). A special script was written to download and convert all the data into jsonlines. During transformation, a unique hash was assigned to each pair, which was later used for a simple deduplication. The resulting dataset was then split into smaller chunks to allow for the parallel processing on the GPUs we had. In addition to the hash used for unique identification, the source column was added to allow us to trace every sentence pair back to the sources where it was found.

After merging and deduplication, we had about 55 million sentence pairs for further evaluation. The total size of the corpus is around 23 gigabytes.

Dataset	Sentences	Deduplicated
CCMatrix	20,240k	19,986k
ParaCrawl	14,079k	13,757k
CCAligned	8,547k	8,113k
MultiMaCoCu	6,406k	5,831k
XLEnt	3,671k	3,392k
OpenSubtitles	10,541k	779k
wikimedia	757k	698k
WikiMatrix	681k	540k
ELRC-5214-A	495k	443k
ELRC-5183-SciPar	306k	301k

Table 1: Top 10 parallel corpora from opus.nlpl.eu ordered by amount of sentences after deduplication, thousands of sentences

3.2 Automatic Evaluation Framework

To apply the quality estimation models, we used the *unbabel-comet* package for the Comet/xCOMET family of metrics and the *metricx* repository for the MetricX family (see Appx. B for the details). For the models available in different sizes and quantization, we picked the largest ones that can fit on available GPUs. We made an exception for the wmt23-cometkiwi-da metric. We used both XL and XXL versions to see if their accuracy differed (see Fig. 2). We also made a comparative analysis on 2 million samples to investigate the Comet model performance under different matmul precision[7] settings. Our finding shows that running the model with medium matmul precision speeds up the evaluation process threefold, while the difference in calculated scores is neglectable (median: 0.000059, mean: 0.000081 on a 0-1 continuous scale). To account for differences in scales used by MetricX and COMET, we applied the following rescaling:

$$metricx_{adj} = 1 - \frac{metricx}{25} \qquad (1)$$

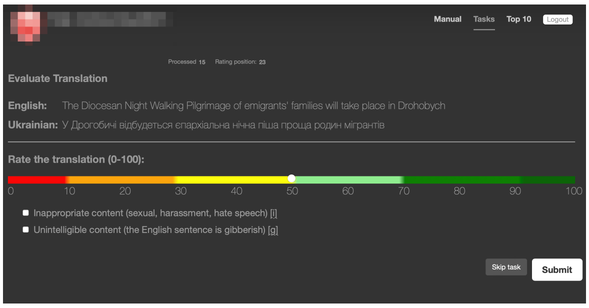
because MetricX has an inverted 0-25 scale.

3.3 Sampling Strategy for Human Evaluation

To sample initial 10,000 pairs for the human evaluation, we stratified the dataset, randomly selecting pairs from the cohorts based on the sentence lengths and assigned average scores of wmt22-cometkiwi-da, wmt23-cometwiki-da-xxl, wmt23-cometwiki-da-xl, and XCOMET-XXL models, which we had already calculated at this point. The cohorts were defined based on the joint decile classification of the two variables. Specifically, the dataset was partitioned into 100 distinct groups by cross-tabulating the deciles of each variable (i.e., 10 deciles × 10 deciles). A representative sample was subsequently drawn by randomly selecting observations from each of these 100 groups. This strategy allowed us to run human evaluations on

[6] https://en.wikipedia.org/wiki/Translation_Memory_eXchange

[7] https://pytorch.org/docs/stable/generated/torch.set_float32_matmul_precision.html



Processed 15 Rating position: 23

Evaluate Translation

English: The Diocesan Night Walking Pilgrimage of emigrants' families will take place in Drohobych

Ukrainian: У Дрогобичі відбудеться єпархіальна нічна піша проща родин мігрантів

Rate the translation (0-100):

- ■ Inappropriate content (sexual, harassment, hate speech) [i]
- ■ Unintelligible content (the English sentence is gibberish) [g]

Figure 1: Crowdsourcing solution for pairs evaluation

sentences of different lengths and quality. We initially aimed to completely evaluate at least 5000 pairs using the resources we found.

3.4 Human Evaluation Protocol

To evaluate the stratified sample, we developed an online crowdsourcing solution using our framework Vulyk[8] (see Fig. 1). This solution allows users to register and score the presented pairs. For the evaluation, we used a pseudo-continuous 0 to 100 scale, mirroring the setup found in (Graham et al., 2013), (Guzmán et al., 2019), which is widely used for Direct Assessment datasets (Graham et al., 2016). In addition to the score, we added two flags so annotators can mark pairs with inappropriate (sexual, harassment, hate speech) or unintelligible content, as we were aware beforehand that some corpora were automatically crawled from the web and may contain such flaws. We also wrote a simple instruction for the grading using the same ranges as found in the original works:

- **0-10**: Incorrect translation
- **11-29**: A few correct keywords, but the meaning is different
- **30-50**: Major mistakes in translation
- **51-69**: Understandable but contains typos or grammatical errors
- **70-90**: Preserves semantics closely
- **91-100**: Perfect translation

Each pair was assigned at random, and to close the task, we required it to have at least three scores from three annotators. During the annotation, we involved more than twenty participants from two different groups of students of linguistic faculties with known proficiency in both English and Ukrainian. The leaderboard was available during the process to encourage students to deliver more evaluations. The final dataset received 9775 evaluated pairs. To ensure the reliability of the results, the scores provided by experts who evaluated fewer than 50 translation pairs were excluded from the final analysis.

3.5 Statistical Analysis Methods

Upon completing the automatic and human evaluation, we did a thorough statistical analysis. It covered both descriptive statistics and inferential methods. We computed standard descriptive statistics for both expert ratings and model scores, including means, standard deviations, and measures of asymmetry. These statistics are provided in Appendix A. The shapes of the distributions, as illustrated in Figure 2, indicate noticeable skewness and asymmetry. Before the further analysis we transformed raw expert scores into percentile ranks to address the non-continuous nature of the data and normalized some of model scores (MetricX23 and MetricX24). We calculated correlation matrices using pairwise complete observations to assess inter-expert agreement and estimated the Intraclass Correlation Coefficient (ICC) using a mean-rating, absolute-agreement, 2-way random-effects model. Finally, we constructed predictive models, including multiple linear and beta regressions using all QE model scores and quadratic regression based on averaged models' scores, to estimate human quality judgments from automatic metrics.

This multi-stage approach provided enough data for analyzing the performance of quality estimation models and their correlation with human judgments, which we present in the following sections.

4 Results

Our analysis of the English-Ukrainian parallel corpus provided some important findings regarding the relationship between automatic quality estimation and human evaluation.

4.1 Descriptive Statistics

The final dataset comprises 9775 translation pairs that received an expert rating. Among the translation pairs, 250 received only one expert ranking, 746 received two rankings, 8528 received three rankings, and 116 received four or five rankings. Notably, only 710 pairs received three rankings from the same set of experts.

Annotators flagged 556 pairs (5.7%) as garbled source text, and 376 pairs (3.8%) were marked as inappropriate or explicit content. Overall, approximately 9.2% of the translation pairs can be

[8] https://github.com/lang-uk/vulyk-translations

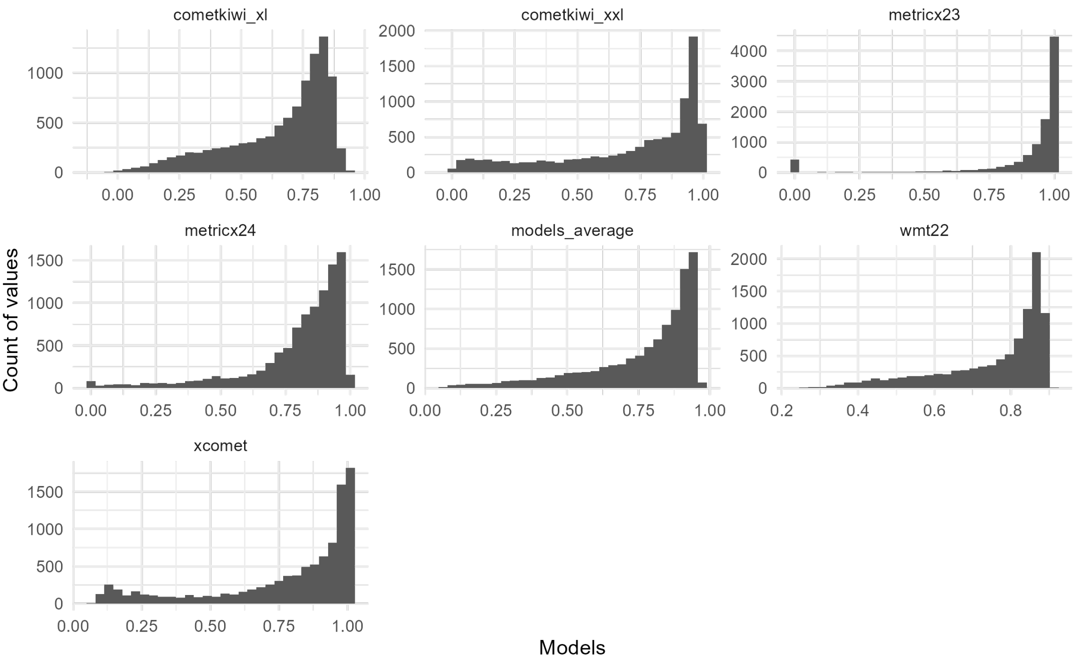


Figure 2: Histograms of model-generated scores and their average

considered invalid for the task due to issues in the dataset.

Human evaluators demonstrated varied scoring patterns, with median scores ranging from 57 to 99 on the 0-100 scale. Most experts who evaluated more pairs (>1000) tended to assign higher scores more frequently, with medians between 67 and 99. This pattern suggests a tendency toward leniency or scoring consistency over time. Contrarily, evaluators who assessed fewer pairs exhibited visible variability in their scoring distributions.

Automatic evaluation models generally assign higher quality scores than human experts. The Google MetricX-24-hybrid-xxl-v2p6 model was quite optimistic with a median score of 0.98 (on the rescaled 0-1 scale), while the wmt23-cometkiwi-da-xl model was the most conservative with a median of 0.73. The wmt22-cometwiki-da model showed the lowest standard deviation (0.14) among all evaluated models, showing better consistency in scoring. For the MetricX models, the histograms exhibit noticeable peaks near zero. This is likely attributable to the nature of the models, which apply linear regression to predict scores and subsequently clip the predicted values outside the 0–25 range. Histograms of the score's distribution can be seen in Fig. 2

4.2 Inter-Annotator Agreement Analysis

We examined the correlation matrix of expert ratings and model-generated scores to assess IAA. Our analysis indicated a higher degree of agreement among QE models, supported by strong correlations.

In contrast, expert ratings showed greater variability, including some cases of strong disagreement between individual evaluators. Given that experts evaluated randomly assigned subsets of translation pairs, we calculated the Intraclass Correlation Coefficient (ICC) based on the ratings from three experts who each evaluated more than 2,000 pairs. Out of these, 710 pairs were evaluated by all three selected experts. Using a mean-rating, absolute-agreement, 2-way random-effects model, we found the level of inter-rater reliability fell within the range of "poor" to "moderate" (ICC = 0.428, 95% CI: 0.252-0.562) (Koo and Li, 2016).

We transformed the raw scores into percentile ranks to address the non-continuous nature of expert ratings despite using a 0-100 scale. This transformation slightly increased the ICC value to 0.542 (95% CI: 0.496-0.585).

4.3 Correlation Between Automatic Metrics and Human Judgments

The ICC calculated for the same set of translation pairs using model scores yielded a slightly

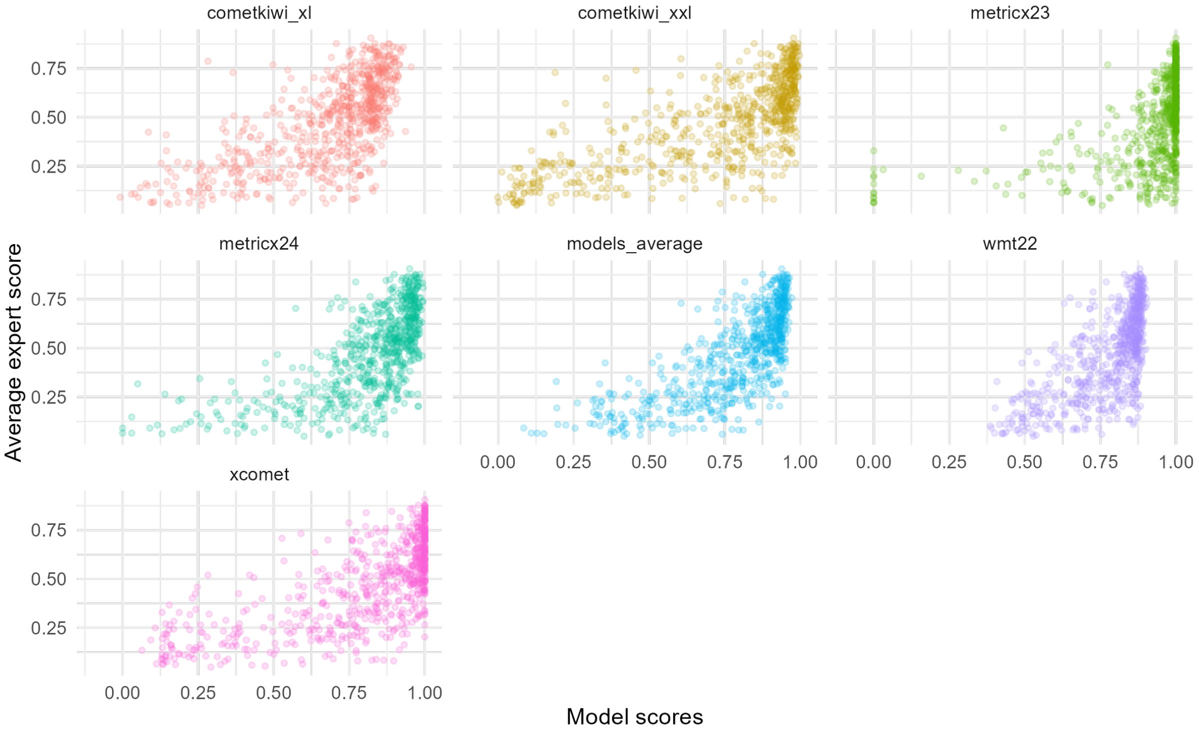


Figure 3: Scatter plot of model scores versus average expert percentile ranks

higher value (ICC = 0.634, 95% CI: 0.538-0.706) compared to the expert ratings. Notably, it was primarily influenced by the Google models. Excluding these models increased the models' ICC to 0.704 (95% CI: 0.635-0.757), indicating moderate reliability that is significantly higher than the ICC observed for the expert ratings. The correlation heatmap (see 4) analysis revealed varying degrees of association between individual models and human evaluations. Models from the same family (COMET or MetricX) tended to correlate more strongly with each other than models from different families. This observation suggests that different model families might be capturing different aspects of translation quality. Correlation patterns can be seen on the scatter plot 3.

4.4 Performance of Regression Models

We constructed three regression models to investigate whether it is possible to predict the human score based on model-generated scores. The first linear model, which incorporated all six model-generated scores to predict the average expert score, explained more than half of the variance (R^2 = 0.559). The most significant contributors to this model were the xcomet, wmt22-cometkiwi-da, and wmt23-cometkiwi-da-xxl models (see Eq. 2).

$$\begin{aligned} score_{linear} = & -0.19600 \\ & + 0.23592 \times \text{xcomet} \\ & + 0.40094 \times \text{wmt\_22} \\ & + 0.18321 \times \text{cometkiwi\_xl} \\ & - 0.02066 \times \text{cometkiwi\_xxl} \\ & - 0.06996 \times \text{metricx23} \\ & + 0.10835 \times \text{metricx24} \end{aligned} \tag{2}$$

Recognizing that building a regression model with correlated variables violates the assumption of multicollinearity, and observing non-linear patterns in the scatter plots, we adopted an alternative approach: averaging the scores from all models and constructing a quadratic regression model. It provided a better fit, explaining 59.2% of the variance. This improvement suggests a non-linear relationship between averaged model-generated scores and expert judgments, observed on the Fig. 3 of model scores versus expert percentile ranks (see Eq. 3).

$$\begin{aligned} score_{quadratic} = & 0.29470 \\ & - 0.87041 * \text{model\_avg} \\ & + 1.33003 * \text{model\_avg}^2 \end{aligned} \tag{3}$$

The non-linear nature of this relationship indicates that automatic quality estimation models may not consistently align with human judgments across

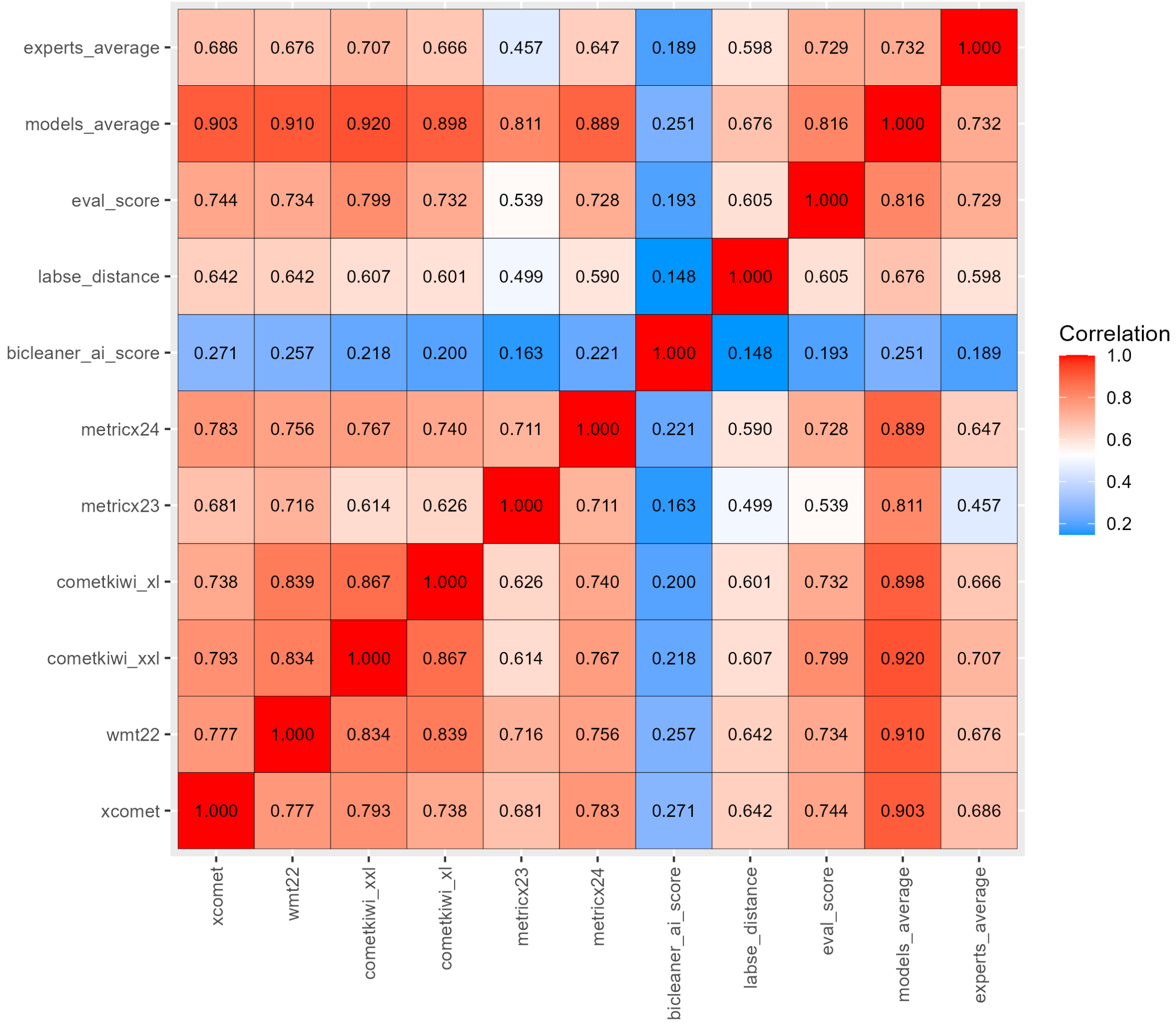


Figure 4: Correlation heatmap between expert average scores and automated metrics. models_average only include 6 QE models

the entire range of translation quality, particularly for translations of moderate quality.

Since the distribution of values was constrained to the interval (0, 1), we applied beta regression to model the proportion of expert scores using model scores as predictors. A logit link function was employed. The model, with estimated coefficients substituted, is specified in Eq. 4.

$$\begin{aligned}\text{logit}(score) = &-3.336\\ &+ 1.046 * \text{xcomet}\\ &+ 1.933 * \text{wmt22}\\ &+ 0.676 * \text{cometkiwi\_xxl}\\ &+ 0.066 * \text{cometkiwi\_xl}\\ &- 0.250 * \text{metricx23}\\ &+ 0.678 * \text{metricx24}\end{aligned} \tag{4}$$

The precision parameter estimate ($\phi = 10.720$) indicates relatively low dispersion around the predicted means. The model demonstrates good fit, with a pseudo R² (McFadden, 1972) of 0.57.

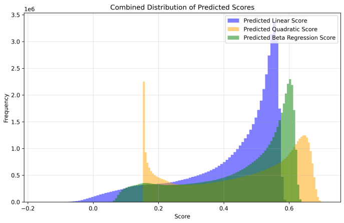


Figure 5: Distribution of predicted scores for three regression models

model	min	max	avg	q1	q2	q3
linear	-0.17	0.64	0.42	0.34	0.48	0.54
quad.	0.15	0.71	0.46	0.31	0.50	0.61
beta reg	0.04	0.68	0.45	0.34	0.51	0.58

Table 2: Characteristics of the trained models calculated on the full dataset

4.5 Final Dataset

In the last step, we applied three models to a whole dataset to calculate the adjusted model scores. Table 2 and Fig. 5 contain the key statistical properties of the score distributions. The threshold for the filtering should be considered according to the task at hand and the amount of data available for a particular language pair and required for model training. As a rule of thumb, for the quadratic model, we recommend:

- A threshold of 0.5 would provide a balanced trade-off between quality and quantity, retaining approximately 50% of the corpus (median score: 0.497).
- A conservative threshold of 0.62 retains only the highest quality pairs (top 20% of the corpus).
- Applications requiring more training data might use 0.31 (retaining 75% of the corpus) to exclude only the clearly problematic pairs.

4.6 Additional Experiments

To cover models and frameworks beyond the QE, we conducted a small set of experiments calculating scores on the human-evaluated dataset using the bicleaner-ai framework and cosine similarity of LaBSE sentence embeddings, calculated for the original and translated text. While bicleaner-ai showed a poor correlation with expert and model average (0.19 and 0.25, respectively), LaBSE cosine similarity produced visibly better results (0.59 and 0.68), which makes it a good candidate for inclusion into the ensemble of models on the subsequent iterations of our experiments. Correlation of these two models to other models and expert average can be seen on the Fig. 4.

We also trained a few additional models, such as XGBoost and SVR, using k-fold validation; however, we observed no improvement over our basic models, so we are not reporting these results.

Additionally, at the very last stage of the research, we conducted a set of experiments on a human-annotated subset of the dataset using the LLM-as-a-Judge method and a detailed prompt (see Appx. C), which asked the model to justify its score. For the commercial model Gemini Pro Preview 2.5, we achieved a correlation of 0.76, and for Gemma 3 27B, 0.73, which places this technique at the top of the leaderboard at the cost of additional compute.

5 Applications

Our research findings can be applied to create a similar evaluation and cleaning pipeline for other language pairs or on newly obtained data for the English-Ukrainian language pair as the number of publicly available corpora and the volume of the data continues to grow. Better filtration of the training data will result in better NMT models, thus bringing us closer to the ultimate task of seamless, high-quality text translation. The insights about the QE models performance might help others reduce the computational complexity of the task by selecting only the best-performing models. The existing methodology for human evaluation is now operationalized into a plugin for a crowdsourcing framework Vulyk[9], making it easy to run similar evaluations or create new Direct Assessment datasets for other languages. The human evaluation dataset can be used to calibrate the QE models further, fit new ensemble models, or assess the quality of other metrics not included in the current research.

Today, we are releasing our framework Vakula[10], which allows users to download, parse, deduplicate, and evaluate the parallel corpora from the Opus Open Parallel Corpora project. We are also releasing a combined and deduplicated corpus of English-Ukrainian parallel sentences with all the scores from QE models and our ensemble mod-

[9] https://github.com/mrgambal/vulyk

[10] https://github.com/lang-uk/vakula

els[11]. Finally, we are publishing the crowdsourcing plugin for human evaluation tasks, the annotator manual, and the raw data obtained from our experiment[12].

6 Conclusion and Future Work

In this paper, we developed a methodology for automatically evaluating and filtering large-scale parallel corpora for NMT. We applied six modern QE models to score 55 million English-Ukrainian sentence pairs and conducted human evaluation on a stratified sample of over 9,775 pairs.

Here are some important findings:

- Automatic QE models showed moderate agreement with human judgments, with our best ensemble model explaining approximately 60% of the variance in averaged expert ratings.
- We found that a quadratic model based on averaged QE scores outperformed linear models, indicating a non-linear relationship between automatic metrics and human quality perception. Akcnowledging the nature of the data distribution, the beta regression can be applied as well.
- QE models demonstrated higher inter-rater agreement than human evaluators, suggesting that while models may not fully capture human judgment, they provide more consistent evaluation than individual annotators.
- The comparative analysis of QE models showed that Unbabel's COMET family and Google's MetricX family have different scoring patterns, with Google models generally assigning higher scores. Our additional experiments demonstrated that simpler models like the LaBSE sentence transformer performed on par with some specialized QE models. This can be handy for pre-filtering or setups with a limited compute.
- Our additional experiments with LLM-as-a-Judge have demonstrated strong performance, on par with the model ensemble, for both Gemma3 27B and Gemini 2.5 Pro Preview.

The evaluated corpus with quality scores allows researchers to select appropriate score thresholds based on their specific needs and input data.

For future work, we plan to:

- Run an additional human evaluation round with professional translators to score at least 1000 pairs with four experts.
- Evaluate the downstream impact of corpus filtering on NMT performance by training models on filtered datasets.
- Perform ablation study on downstream task, training NMT models using data, filtered under different thresholds.

By releasing our framework, evaluated corpus, and human evaluation data, we hope to facilitate further research in parallel corpus filtering and quality estimation and ultimately contribute to higher-quality neural machine translation systems.

7 Acknowledgments

We express our gratitude to Mariia Shvedova, Halyna Yarotska, Anna Pospekhova, and Mariana Romanyshyn.

Students of Applied Linguistics at I. Mechnikov Odesa National University: Anna Drobotova, Alina Onishchenko, Daryna Kholomieieva, Anna Polieshchuk, Anna Lovochkina, Sofiia Kondratiuk, Yelyzaveta Petrova, Anastasiia Mikheieva.

Participants of the Corpus Linguistics online course (as part of the digital philology program for Ukrainian students, supported by the University of Jena and the DAAD Foundation): Olena Romaniuk, Yurii Petrov, Veronika Moroz, Olha Tochylina, Alina Movchan, Zhanna Voloshko, Olena Bezverkhna, Mariia Morozova, Daryna Moroz, Anna Khuhaieva, Sofiia-Tereza Onysko, Kateryna-Olha Vozniuk, Nataliia Sheremet, Maksym Vodianko

We also want to thank the Talents for the Ukraine project of the Kyiv School of Economics and ELEKS for the grants on compute resources.

Limitations

We acknowledge the following limitations of the work done in this paper:

- All three regression models were developed using a relatively small subsample of data and expert rankings characterized by moderate inter-expert agreement. As a result, the predicted expert ranks exhibit a limited range and do not approach the extreme values of 0 or 1.
- Using students of linguistics rather than professional translators might affect the quality and variability of the evaluation.

[11] https://huggingface.co/datasets/lang-uk/FiftyFiveShades

[12] https://github.com/lang-uk/vulyk-translations

- The work focuses on a particular language pair, and similar research might yield different results for other language pairs.
- The findings of this paper have yet to be confirmed by extrinsic evaluation.
- The quality of the corpora we used and their domains is beyond our control.

The authors acknowledge using Grammarly for paraphrasing and revision in the process of writing this paper and Github Copilot autocomplete when working on the code.

References

Yvette Graham, Timothy Baldwin, Meghan Dowling, Maria Eskevich, Teresa Lynn, and Lamia Tounsi. 2016. Is all that glitters in machine translation quality estimation really gold? In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3124–3134, Osaka, Japan. The COLING 2016 Organizing Committee.

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.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Md. Arid Hasan, Prerona Tarannum, Krishno Dey, Imran Razzak, and Usman Naseem. 2024. Do large language models speak all languages equally? a comparative study in low-resource settings. *Preprint*, arXiv:2408.02237.

Marcin Junczys-Dowmunt. 2018. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 888–895, Belgium, Brussels. Association for Computational Linguistics.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *Preprint*, arXiv:2001.08361.

Terry K. Koo and Mae Y. Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2):155–163.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: A multilingual and document-level large audited dataset. *Preprint*, arXiv:2309.04662.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

Daniel McFadden. 1972. Conditional logit analysis of qualitative choice behavior.

Yurii Paniv, Dmytro Chaplynskyi, Nikita Trynus, and Volodymyr Kyrylov. 2024. Setting up the data printer with improved English to Ukrainian machine translation. In *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024*, pages 41–50, Torino, Italia. ELRA and ICCL.

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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. 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.

Víctor M. Sánchez-Cartagena, Marta Bañón, Sergio Ortiz-Rojas, and Gema Ramírez. 2018. Prompsit‘s submission to WMT 2018 parallel corpus filtering shared task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 955–962, Belgium, Brussels. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Jörg Tiedemann. 2016. OPUS – parallel corpora for everyone. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation: Projects/Products*, Riga, Latvia. Baltic Journal of Modern Computing.

Jaume Zaragoza-Bernabeu, Gema Ramírez-Sánchez, Marta Bañón, and Sergio Ortiz Rojas. 2022. Bicleaner AI: Bicleaner goes neural. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 824–831, Marseille, France. European Language Resources Association.

Tianyang Zhong, Zhenyuan Yang, Zhengliang Liu, Ruidong Zhang, Yiheng Liu, Haiyang Sun, Yi Pan, Yiwei Li, Yifan Zhou, Hanqi Jiang, Junhao Chen, and Tianming Liu. 2024. Opportunities and challenges of large language models for low-resource languages in humanities research. *Preprint*, arXiv:2412.04497.

A Statistics

Table 3 contains descriptive statistics on the 6 QE models. Table 4 contains descriptive statistics on annotators.

B Models

Table 5 contains the information on the used Quality Estimation models, their backbone models and number of parameters.

C Prompts

Your task is to evaluate the quality of a translation from English to Ukrainian. Carefully read both the English and Ukrainian sentences and assign a score based on the accuracy of the translation. Rate the translation on a scale of 0 to 100, where:

0-10: INCORRECT TRANSLATION

- The translation is completely wrong or incomprehensible - No meaningful connection to the original English text - May be gibberish, unrelated content, or severely corrupted text - Ukrainian readers would have no idea what the original English meant - Examples: wrong language, scrambled words, completely different meaning

11-29: FEW CORRECT KEYWORDS, MEANING IS DIFFERENT

- Only a few individual words are correctly translated - The overall meaning is significantly different from the original - Key concepts, actions, or subjects are mistranslated - Ukrainian readers would understand some words but get the wrong message - The translation might be partially readable but conveys incorrect information - Missing critical information or contains major factual errors

30-50: MAJOR MISTAKES IN TRANSLATION

- The general topic or domain is recognizable but with serious errors - Multiple important words or phrases are incorrectly translated - Sentence structure may be broken or very awkward - Some key information is preserved but significant details are wrong - Ukrainian readers can guess the general topic but many specifics are unclear - May include incorrect technical terms, wrong numbers, or misidentified entities - Grammar errors that significantly impact meaning

51-69: UNDERSTANDABLE BUT CONTAINS ERRORS

- The main meaning is generally preserved and understandable - Contains noticeable typos, grammatical errors, or awkward phrasing - Minor mistranslations that don't completely change the meaning - Word order issues or unnatural Ukrainian sentence structure - Ukrainian readers can understand the message despite the errors - May have inconsistent terminology or slightly incorrect word choices - Punctuation or capitalization errors that affect readability

70-90: PRESERVES SEMANTICS CLOSELY

- Accurately conveys the original meaning with minor imperfections - Natural Ukrainian grammar and sentence structure - Appropriate word choices and terminology - May have very minor stylistic issues or slightly awkward phrasing - All key information is correctly translated - Ukrainian readers can easily understand without confusion - Demonstrates good understanding of both languages

91-100: PERFECT TRANSLATION

- Flawless translation that perfectly captures the original meaning - Natural, fluent Ukrainian that sounds native - Appropriate style and register for the context - All nuances, tone, and subtleties are preserved - Perfect grammar, spelling, and punctuation - Reads as if originally written in Ukrainian - No improvements needed

When evaluating, consider: 1. Accuracy of meaning and content 2. Grammar and syntax correctness 3. Natural flow and readability in Ukrainian 4. Completeness (nothing important omitted or added) 5. Appropriate word choices and terminology

Please provide the reason first, followed by a score. Format your evaluation in the JSON structure below: {"reason": "reason for the score", "score": int}

	n	mean	std	mdn	trmd	mad	min	max	rng	skew	kurt	se
xcomet	9775	0.78	0.27	0.89	0.83	0.16	0.05	1	0.95	-1.28	0.39	0.003
wmt22	9775	0.76	0.14	0.82	0.78	0.08	0.23	0.90	0.68	-1.25	0.60	0.001
cometkiwi_xxl	9775	0.71	0.29	0.82	0.75	0.21	-0.03	1	1.03	-0.98	-0.30	0.003
cometkiwi_xl	9775	0.66	0.21	0.73	0.68	0.16	-0.10	0.95	1.05	-1.04	0.18	0.002
metricx23	9775	0.89	0.23	0.98	0.95	0.03	0	1	1	-2.91	7.77	0.002
metricx24	9775	0.79	0.20	0.86	0.83	0.12	0	1	1	-1.89	3.46	0.002
models_average	9775	0.76	0.20	0.84	0.80	0.13	0.06	0.97	0.91	-1.37	1.18	0.002

Table 3: Descriptive statistics for the scores assigned by the automatic evaluation models

	n	mean	std	mdn	trmd	mad	min	max	rng	skew	kurt	se
expert_1	620	75.27	29.19	89.00	80.23	16.31	0	100	100	-1.24	0.34	1.17
expert_2	501	79.94	22.22	85.00	84.29	14.83	0	100	100	-1.77	2.93	0.99
expert_3	5392	72.60	21.45	77.00	75.56	16.31	0	100	100	-1.41	2.11	0.29
expert_4	480	58.64	30.59	60.00	59.72	43.74	0	100	100	-0.17	-1.33	1.40
expert_5	551	60.24	29.46	68.00	62.51	29.65	0	100	100	-0.59	-0.88	1.25
expert_7	461	91.07	17.74	98.00	95.40	2.97	0	100	100	-3.47	12.71	0.83
expert_8	5425	85.67	26.60	99.00	92.30	1.48	0	100	100	-1.94	2.61	0.36
expert_11	495	87.44	11.89	90.00	89.57	2.97	6	100	94	-4.21	21.83	0.53
expert_12	3124	70.31	33.31	87.00	75.37	17.79	0	100	100	-1.13	-0.18	0.60
expert_13	2151	60.73	26.92	67.00	63.30	26.69	0	98	98	-0.70	-0.46	0.58
expert_14	363	96.89	1.84	97.00	96.97	1.48	91	100	9	-0.48	-0.10	0.10
expert_15	331	77.10	28.96	89.00	83.07	13.34	0	100	100	-1.53	1.18	1.59
expert_16	293	56.34	31.21	59.00	57.38	44.48	0	100	100	-0.17	-1.45	1.82
expert_18	307	53.35	26.48	57.00	54.87	25.20	0	96	96	-0.48	-0.68	1.51
expert_19	2136	78.88	23.88	88.00	83.96	11.86	0	100	100	-1.77	2.39	0.52
expert_20	310	68.71	33.17	86.00	73.50	16.31	0	100	100	-1.08	-0.31	1.88
expert_22	2653	62.80	34.71	72.00	65.65	40.03	0	100	100	-0.46	-1.28	0.67
expert_23	282	81.88	25.46	95.00	87.57	7.41	0	100	100	-1.84	2.77	1.52
expert_24	300	62.87	32.89	71.00	65.85	34.84	0	100	100	-0.62	-0.96	1.90
expert_26	300	74.20	25.36	85.00	78.14	15.57	0	100	100	-1.17	0.29	1.46
expert_27	345	63.18	33.25	72.00	65.95	37.07	0	100	100	-0.53	-1.17	1.79
expert_29	297	61.00	30.62	58.00	63.67	28.17	0	100	100	-0.48	-0.44	1.78
expert_30	302	73.76	33.55	90.00	79.66	14.83	0	100	100	-1.26	0.20	1.93
expert_32	323	65.45	40.05	90.00	69.26	14.83	0	100	100	-0.69	-1.30	2.23

Table 4: Descriptive statistics for the scores assigned by the annotators

Abbreviation	Family	HuggingFace model handle	Base model	Params
cometkiwi_xxl	CometKiwi	Unbabel/wmt23-cometkiwi-da-xxl	XLM-R-XXL	10.5B
cometkiwi_xl	CometKiwi	Unbabel/wmt23-cometkiwi-da-xl	XLM-R-XL	3.5B
metricx24	MetricX	google/metricx-24-hybrid-xxl-v2p6-bfloat16	mT5-XXL	13B
metricx23	MetricX	google/metricx-23-qe-xxl-v2p0	mT5-XXL	13B
xcomet	XComet	Unbabel/XCOMET-XXL	XLM-R-XXL	10.7B
wmt22	CometKiwi	Unbabel/wmt22-cometkiwi-da	InfoXLM	n/a

Table 5: Detailed information on used QE models