Optimizing Quality Estimation for Low-Resource Language Translations: Exploring the Role of Language Relatedness

Archchana Sindhujan*[1[0000−0002−6467−6873]], Diptesh Kanojia[1[0000−0001−8814−0080]], and Constantin Orăsan[1[0000−0003−2067−8890]]

University of Surrey, UK
{a.sindhujan,d.kanojia,c.orasan}@surrey.ac.uk

Abstract. Evaluation of machine translation (MT) is vital to determine the effectiveness of MT systems. This paper investigates quality estimation (QE) for machine translation (MT) for low-resource Indic languages. We analyse the influence of language relatedness within linguistic families and integrate various pre-trained encoders within the MonoTransQuest(MonoTQ) framework. This entails assessing models in single-language configurations before scaling up to multiple-language setups, focusing on languages within and across families, and using approaches grounded in transfer learning. Experimental outcomes and analyses indicate that language-relatedness significantly improves QE performance over baseline, sometimes even surpassing state-of-the-art approaches. Across monolingual and multilingual configurations, we discuss strategic encoder usage as a simple measure to exploit the language interactions within these models improving baseline QE efficiency for quality estimation. This investigation underscores the potential of tailored pre-trained encoders to improve QE performance and discusses the limitations of QE approaches for low-resource scenarios.

Keywords: multilingual · pre-trained encoders · efficiency

1 Introduction

Quality estimation seeks to evaluate the reliability of translation outputs in the absence of comparative reference texts [28]. The motivation to explore quality estimation in machine-translated content stems largely from the growing need for accurate and dependable machine translation in various linguistic and cultural contexts.

The motivation for this study stems from the computational challenges posed by Quality Estimation (QE) models for machine translation, particularly in low-resource settings. These models often depend on large and complex models, leading to high costs and latency. This research aims to enhance the efficiency and effectiveness of QE by simply optimizing encoder configurations and leveraging language relatedness, focusing on resource-scarce Indic languages. The core investigation centres on whether the interaction among different

Proceedings of New Trends in Translation and Technology, pages 170–190
3–6 July, 2024, Varna, Bulgaria. ©NeTTT 2024
https://doi.org/10.26615/issn.2815-4711.2024_014

languages and language families enhances or compromises the accuracy of quality estimations, both individually and collectively. Various experiments are carried out to discover insights into the impact of interactions between languages within QE models.
The main contributions of this paper are,

- Our analysis reveals that the linguistic similarities present within the languages improve the efficiency of pre-trained encoder-based QE models.
- We identified that the MonoTQ-InfoXLM-large encoder outperforms other pre-trained encoders in the majority of experimental settings, suggesting that the contrastive learning approach effectively supports Quality Estimation tasks.

The organization of this paper includes Section 2 providing essential background on quality estimation, Section 3 detailing the methodological framework and experimental settings, Section 4 presenting the results, and insights, and followed by Section 5 which delves into error analysis. The concluding section summarizes key findings, draws conclusions, and proposes directions for future research endeavours.

2 Background

In recent years, the domain of quality estimation has experienced significant advancements. The evolution of QE models began with models based on feature engineering [29, 27]. It then progressed through various stages, from basic neural networks [14] to more complex deep neural network architectures [10, 13], and has now reached the phase of utilizing Large Language Models [15].

Currently, transformer-based approaches are at the forefront of QE techniques [4, 23, 21, 20, 19, 1]. There's also an emerging trend of employing ensemble methods, where various QE models are integrated to yield a more consistent assessment of MT quality [12, 13, 2, 18, 3, 8, 23, 1].

Top-performing systems from the WMT23 Quality Estimation shared tasks [4], representing the current state-of-the-art, predominantly utilize ensemble methods. For the En-Mr and En-Gu language pairs, the Unbabel-IST's cometkiwi [22] with XLM-R XL and XXL pre-trained encoders, achieved the best results. Meanwhile, for the En-Hi, En-Ta, and En-Te, the HW-TSC's approach of combining multi-lingual encoders (XLMRoBERTa, InfoXLM, RemBERT) with a task-specific downstream layer showed the best performance [30, 4].

These methods rely on neural networks that utilize a significant amount of computational power and disk space usage. These characteristics of current QE systems limit their widespread adoption in practical scenarios. This leads us to investigate how to improve the efficiency of the models by examining language-relatedness among different language pairs within various settings with comparatively smaller models which require lesser computational power and disk space. The TransQuest framework [21] was selected for this study due to its

adaptable architecture, allowing flexibility in modifying the backend encoder and convenient adjustment of hyperparameter configurations.

In our research, we utilized the Direct Assessment (DA) score datasets from the WMT23 Quality Estimation shared tasks [4], as shown in Appendix A, where scores range from 0 (indicating the lowest quality) to 100 (indicating a perfect translation). These scores are provided by human evaluators and are based on comprehensive guidelines that include criteria like adequacy, fluency, and overall accuracy of the translation which aligns with FLORES methodology outlined by Guzmán et al. (2019) [9]. Each translation is evaluated by at least three different annotators, and these scores are averaged and standardized using z-scores to form a unified quality metric known as the z-mean.

3 Methodological Framework

3.1 Architecture

Our methodology leverages the approach outlined by TransQuest [21], where the input to the model is structured by concatenating the source sentence with its corresponding translated version, separated by a [SEP] token to differentiate the source from the translation. The [CLS] token is positioned as the initial token in the sequence, which facilitates the assimilation of contextual cues from all other tokens in the sequence via the multi-head-attention mechanism [31]. Subsequently, each word in the sequence is assigned its individual embedding, enabling the model to capture nuanced semantic information and relationships within the input text. For various experimental setups, we employ distinct pre-trained encoders (see section 3.2).

In the final layer of the architecture, the softmax function is employed to predict the quality score. During training, optimization was guided by minimizing the Mean-Squared Error (MSE) as the loss function.

3.2 Pre-trained Models

Based on the architecture described in section 3.1 our study incorporates three Pre-trained Language Models (PTLMs): XLMR-large, XLM-V, and InfoXLM-large.

XLM-Roberta (XLMR) : This model builds on the XLM framework, utilizing cross-lingual pre-training on CommonCrawl dataset [32] to capture nuances across languages, resulting in enriched embeddings [7]. The effectiveness of this model in cross-lingual tasks is enhanced by Masked Language Modeling (MLM) training objectives [7]. MLM strategy predicts masked tokens using context from unmasked tokens in multilingual text from 100 languages. It employs subword tokenization and balanced language sampling without language-specific embeddings, efficiently enhancing cross-lingual learning on a large scale.

XLM-V - Utilizes one-million-token vocabulary, mirroring the XLM-R's dataset, but enhances linguistic diversity representation through judicious vocabulary allocation. The construction of multilingual vocabularies involves training individual SentencePiece models for each language using the Unigram Language Model algorithm, followed by creating lexical representation vectors. These vectors are clustered using the K-Means algorithm to optimize vocabulary capacities effectively. This comprehensive process results in a unified multilingual vocabulary that enhances linguistic accuracy and minimizes token overlap which ensures semantically richer and more intuitive tokenizations than XLM-R, enhancing the ability of the model to understand language-specific nuances correctly across multiple languages [17].

InfoXLM-large - This approach refines the XLM-R framework with information theory, enhancing cross-lingual understanding through mutual information and contrastive learning for semantic alignment. Contrastive learning enhances the alignment between encoded representations of "positive" pairs— correctly translated bilingual sentences and increases the divergence for "negative" pairs, which are not correct translations. This approach effectively enhances the model's ability to distinguish and encode linguistic nuances across languages, increasing the model's multilingual effectiveness [5].

The three chosen pre-trained models are cross-lingual, each with unique and powerful pre-training techniques, facilitating varied strategies for achieving the Quality Estimation (QE) objective. Also combining these models within the architecture explained in 3.1, requires considerably less disk space and shows competitive performance, underscoring the cost-effectiveness of the approach compared to the best-performing systems of WMT23 QE shared tasks [4, 25] as shown in appendix B. Our research aims to determine which pre-training strategies yield the best performance under various experimental configurations and how effectively these models utilize the language-relatedness among training datasets to enhance efficiency in Quality Estimation.

3.3 Experimental Setting

Our experimental methodology is constructed to investigate how the dynamics of multiple languages impact the performance of QE models with different pre-training approaches with a consistent setup of hyperparameters. All the experiments followed a batch size of 8, the use of the Adam optimizer, and a constant learning rate of 2e-5. To optimize the training process, an early stopping criterion was implemented, terminating the training phase upon the absence of improvement in the evaluation loss after a sequence of ten assessments.

For the initial set of experiments we have utilised the following set of low-resourced Indic language datasets of DA score from the WMT23 QE shared task [4]: English-Gujarati (En-Gu), English-Hindi (En-Hi), English-Marathi (En-Mr), English-Tamil (En-Ta), and English-Telugu (En-Te). The MonoTransQuest (MonoTQ) architecture, detailed in section 3.1, served as the foundation for

Table 1. Pearson (r) and Spearman (ρ) correlation scores for different models in two different settings (Multilingual and Monolingual) where Indic language datasets for DA scores are used for training and testing. The highest Pearson correlation scores obtained for each language pair in each setting are marked in bold. The overall best Pearson score among both settings for each language pair is underlined

	En-Gu		En-Hi		En-Mr		En-Ta		En-Te	
Model	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
I. Indic-language-only Multilingual Setting										
MonoTQ-XLMR-large	0.300	0.438	0.430	0.440	-0.117	0.395	0.454	0.482	0.211	0.345
MonoTQ-InfoXLM-large	**0.656**	0.713	**0.726**	0.624	0.030	0.470	**0.662**	0.726	**0.719**	0.462
MonoTQ-XLMV	0.536	0.673	0.687	0.572	**0.426**	0.642	0.559	0.670	0.642	0.464
II. Monolingual Setting										
MonoTQ-XLMR-large	0.391	0.383	**0.536**	0.497	0.076	0.112	**0.558**	0.532	0.286	0.329
MonoTQ-InfoXLM-large	**0.690**	0.653	0.134	0.119	0.508	0.629	0.268	0.303	0.079	0.087
MonoTQ-XLMV	0.595	0.548	0.499	0.438	**0.512**	0.598	0.536	0.497	**0.299**	0.358
WMT23	0.714	0.745	0.644	0.720	0.704	0.735	0.775	0.778	0.394	0.350

constructing experimental models MonoTQ-XLMR-large, MonoTQ-InfoXLM-large, and MonoTQ-XLMV utilizing various pre-trained models outlined in section 3.2.

We utilize Spearman's Rank Correlation [24] as the primary metric , with Pearson Correlation Coefficient [6] serving as the secondary metric to evaluate the model performance.

Experiment 1 - Multilingual and Monolingual Settings : As shown in table 1, this experiment was divided into multilingual and monolingual setups to explore how these settings, along with various pre-trained encoders and Indic language pairs, affect QE. Indic-language-only Multilingual Setting (I) is where all the Indic language pairs are trained together and tested separately. Multilingual configuration evaluates the adaptability and generalizability of encoders across multiple Indic languages, highlighting how collective training with region-specific low-resourced languages influences individual language performance for QE. Monolingual Setting (II) is where each language pair is trained and subsequently evaluated within the same language pair to identify language-specific characteristics and encoder efficiencies for specific low-resourced languages.

Experiment 2 and 3 - Language Family-Specific vs Transfer Learning Approach : Experiment 2 is designed to ascertain the influence of linguistic similarities within closely related language groups, specifically among Indo-Aryan languages (Hindi, Gujarati, Marathi) and Dravidian languages (Tamil, Telugu). By analyzing these groups, the study aims to determine how linguistic proximity within each family affects Quality Estimation (QE) outcomes, as detailed under row I of Table 2. Experiment 3 evaluates the efficacy of transfer learning between the two distinct language families by implementing a cross-testing methodology. This involves training models on Indo-Aryan languages and

Table 2. Pearson (r) and Spearman (ρ) correlation scores for different models in different settings of Experiment 2 & 3. Setting I - Trained and tested with the same language group: Indo-Aryan (IA), Dravidian (Dr). Setting II - Cross-tested within language groups. The highest Pearson correlation scores obtained for each language pair in each setting are marked in bold. The overall best Pearson score among both settings for each language pair is underlined.

Model	En-Gu		En-Hi		En-Mr		En-Ta		En-Te	
	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
I	IA (train) -> IA (test)						Dr (train) ->Dr (test)			
MonoTQ-XLMR-large	0.636	0.591	0.590	0.471	0.487	0.565	-0.056	-0.056	0.066	0.076
MonoTQ-InfoXLM-large	**0.696**	0.655	**0.648**	0.540	0.457	0.616	0.047	0.027	-0.008	-0.024
MonoTQ-XLMV	0.649	0.585	0.617	0.491	**0.529**	0.580	**0.552**	0.515	**0.273**	0.319
II	Dr (train) -> IA (test)						IA (train) -> Dr (test)			
MonoTQ-XLMR-large	-0.030	-0.018	-0.114	-0.035	-0.030	0.417	0.417	0.447	**0.205**	0.236
MonoTQ-InfoXLM-large	0.076	0.075	0.014	0.028	0.098	0.106	**0.553**	0.493	0.196	0.229
MonoTQ-XLMV	**0.269**	0.253	**0.282**	0.295	**0.298**	0.311	0.417	0.447	0.202	0.227

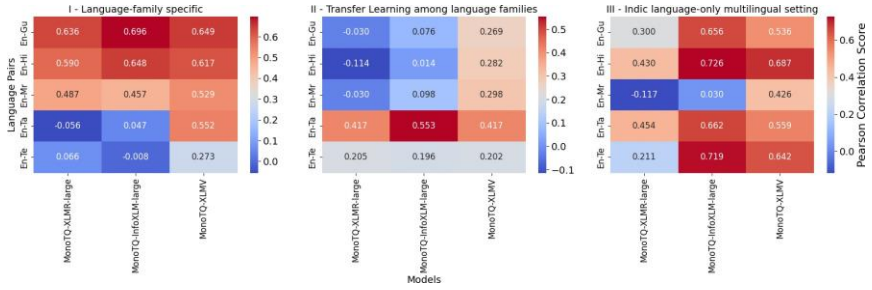

Fig. 1. Heatmap of the Pearson correlation scores from Experiment 2 (I - Language family specific) & Experiment 3 (II - Transfer learning among language families) compared with Experiment 1 (III - Indic languages only multilingual setting)

testing on Dravidian languages, and vice versa, to understand the cross-lingual transfer capabilities of the QE models across linguistically diverse groups. The configurations and results of this testing are presented under row II of Table 2. This approach helps to highlight potential challenges and advantages in applying transfer learning techniques across different language families.

Experiment 4 - Incorporating linguistically diverse non-Indic languages: This experiment assesses how incorporating non-Indic language pairs into the training corpus affects model performance, structured in two settings. The objective is to examine the impact on the performance of quality estimation models by incorporating linguistically diverse datasets with low-resource Indic languages. As shown in table 3 under row I (Setting I), all the DA score datasets which have English on the Source side of the translation are incorporated for the training of the models. This experiment utilizes the WMT DA score datasets (see Appendix A) with English in the source side of the translations, expanding the focus from Indic language pairs to include En-Zh and En-De in the training data. It has been evident that translation models

Table 3. Pearson (r) and Spearman (ρ) correlation scores for different models in different settings (I, II). Setting I -> Incorporating datasets with English in the source side of the translation, Setting II -> Incorporating all the DA score datasets. The highest Pearson correlation scores obtained for each language pair in each setting are marked in bold. The overall best Pearson score among both settings is underlined

	En-Gu		En-Hi		En-Mr		En-Ta		En-Te	
Model	*r*	ρ	*r*	ρ	*r*	ρ	*r*	ρ	*r*	ρ
I. Incorporating English as the Source Language datasets										
MonoTQ-XLMR-large	0.399	0.390	0.371	0.365	0.126	0.319	0.332	0.391	0.197	0.223
MonoTQ-InfoXLM-large	0.660	0.619	**0.625**	0.482	0.426	0.570	**0.707**	0.644	**<u>0.358</u>**	0.356
MonoTQ-XLMV	**0.663**	0.614	0.601	0.475	**0.433**	0.559	0.703	0.636	0.350	0.366
II. Incorporating all the DA score datasets										
MonoTQ-XLMR-large	0.615	0.663	0.618	0.509	0.406	0.576	0.691	0.626	0.349	0.347
MonoTQ-InfoXLM-large	**<u>0.700</u>**	0.657	**<u>0.667</u>**	0.526	**<u>0.448</u>**	0.605	**<u>0.727</u>**	0.685	**0.348**	0.324
MonoTQ-XLMV	0.656	0.605	0.606	0.496	0.462	0.569	0.669	0.633	0.347	0.370

show significantly better performance with English on the source side [11, 16], therefore, this study aims to explore if similar patterns are observed within the Quality Estimation domain. In addition to the language pairs mentioned in the above Setting I, we added 4 additional language pairs (Ru-En, Ro-En, Ne-En, Si-En) for the training of Setting II (see table 3 under row II). This experimental setting examines how data augmentation and increased linguistic diversity (including languages outside their regional languages) impact the efficacy of low-resource language QE models.

Also, we conducted some additional experiments with zero-shot, few-shot, and full-data scenarios to investigate the sample efficiency which can be seen in appendix D.

4 Result and Discussion

The results of experiment 1 as detailed in Table 1, reveal differential performance across Indic-language-only multilingual (I) and monolingual (II) settings. We have added the state-of-the-art performance score from the WMT23 QE shared task as the last row of the table. The optimal results achieved for each language pair with an Indic-language-only multilingual setting closely align with, and in the case of the En-Te language pair, even surpass the current state-of-the-art results from WMT23.

MonoTQ-XLMR-large excels in monolingual contexts, achieving higher Pearson correlation scores compared to its multilingual setup, thus highlighting its proficiency when dedicated to a single language. On the other hand, MonoTQ-InfoXLM-large and MonoTQ-XLMV generally perform better with the primary metric of Pearson correlation scores in multilingual contexts, with the exception of the En-Gu and En-Mr language pairs. For Spearman correlation scores, all three models demonstrate improved performance in multilingual environments across the majority of language pairs.

In general, the Indic-language-only multilingual setting tends to be beneficial for the models, where MonoTQ-InfoXLM-large exhibits the most notable gains. Increased performance in multilingual settings has been seen in past years of WMT QE shared task results too [4, 33, 26]. This suggests that exposure to multiple languages rather than the distinct language, helps models better generalize and capture the nuances of different languages.

In the language-family specific setting (I), as shown in table 2, MonoTQ-InfoXLM-large proved the most robust model across the Indo-Aryan language pairs, achieving the highest scores in correlation. MonoTQ-XLMV shows notable strength in the Dravidian language pairs (En-Ta and En-Te). This suggests that MonoTQ-XLMV is more attuned to the characteristics of Dravidian languages. Also, MonoTQ-XLMV appears to be the most consistent across both language families. From the result of the transfer-learning among the language family configurations, as shown in setting II of the same table, MonoTQ-XLMV was identified as the model with the highest performance in cross-testing from Dravidian to Indo-Aryan languages and with no specific model excelling in the reverse direction.

The heatmap in figure 1 reveals that, although the Indic language-only multilingual setting (III) from Experiment 1 with MonoTQ-InfoXLM-large model shows the best performance, the language family-focused setting (I) yields consistently strong performance throughout the experiments across all three models except En-Te language pair. These results suggest that tailored training on specific language families could yield an acceptable performance with any pre-trained model compared to a generalized or transfer learning approach among most of the low-resourced Indic languages.

From the results obtained from experiment 4 as shown in Table 3 , we can observe that having augmented data with English on the source side did not improve the overall performance of the QE models compared to Setting II. This evidence suggests that the presence of English on the source side does not significantly influence the quality estimation domain for the majority of Indic language pairs, despite its noted impact in translation domains [11, 16]. Setting II from experiment 4 (table 3) shows overall good performance with all the models and all the language pairs except En-Te.

Training augmented with non-Indic languages demonstrates optimal performance for 3 language pairs (En-Gu, En-Mr, En-Ta) with Pearson correlation scores compared to the Indic-language-only configuration (Experiment 1-Setting I) (See appendix E). This evidence suggests that augmenting the dataset with non-Indic languages enhances performance. On the other hand, if we consider the Spearman correlation the integration of Indic languages consistently yields the best outcomes for the majority of the language pairs (except En-Mr), even with comparatively fewer data (See appendix E). This substantiates the hypothesis that language-relatedness among Indic language pairs significantly influences the models' performance scores, surpassing the effects of data augmentation.

Considering all the experiments, MonoTQ-InfoXLM-large performs significantly better than the other two models, suggesting it is the most robust model for most of the language pairs and training configurations (See appendix F). By maximizing mutual information for enhanced cross-lingual understanding and applying contrastive learning to improve semantic alignment between translated pairs, InfoXLM-large excels in capturing and encoding linguistic nuances across multiple languages. This approach significantly enhances its effectiveness in multilingual quality estimation.

Also, we identified that XLM-V is particularly effective in language-family-specific settings (Experiment II) due to its advanced vocabulary management. XLMV optimizes linguistic diversity through a unified vocabulary that enhances linguistic accuracy and minimizes token overlap for each language, making it highly effective for language-specific multilingual representation.

5 Error Analysis

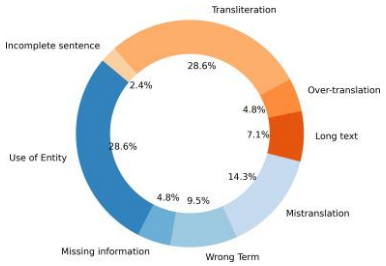

Fig. 2. Error distribution of English-Tamil translation quality estimation models

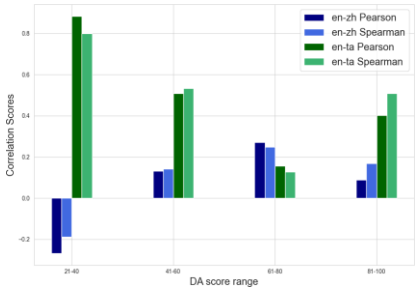

Fig. 3. This figure illustrates the correlation scores across different buckets for both high-resource and low-resource language pairs, providing a detailed comparative analysis.

An error analysis on the English-Tamil dataset was conducted to gain deeper insights into the models' strengths and limitations. The selection of this language pair was primarily due to the availability of native speakers of the target language, who are also fluent in English, to validate any errors. Results obtained from Indic-language-only multilingual setting (see section 3.3 - Experiment 1) were selected for this analysis due to the highest performance in this model configuration compared to other settings.

Predicted z-scores from each model were normalized and then subjected to min-max scaling based on the actual z-score range. Following the normalisation, the top 10% of sentences that exhibited the largest deviation between the predicted and actual scores were identified and extracted from each single-model (MonoTQ-XLMR-large, MonoTQ-InfoXLM-large and MonoTQ-XLMV)

for further analysis. We filtered the common sentences within the top 10% error margin across all three models and analyzed them to identify common factors contributing to inaccuracies. About 50% of the sentences, which ranked with the highest deviation in the scores, exhibited similarities across all three models. These sentences were subjected to further analysis to ascertain if specific types of errors in the input were responsible for the substantial discrepancies between the predicted and the human-annotated scores across these models.

The analysis indicated that a significant proportion of the errors stemmed from the usage of named entities in the source sentence and the transliteration. Sentences involving named entities within the source text, even with correct translations often result in deviations between the predicted quality scores and those annotated by humans. Similarly, the presence of transliteration within a translated sentence frequently leads to discrepancies. The rest of the errors are caused by mistranslation, over-translation, wrong terms in the translation, lengthy texts and incomplete translated sentences. Figure 2 presents the distribution of error categories along with their corresponding percentages identified in the analysis. We have listed some examples of translations in the Appendix H, that led to inaccuracies in the quality estimation model predictions.

The errors involving named entities and transliterations are challenging for QE models because the named entities often require precise translation and preservation across languages to maintain the semantic integrity of the sentence. Wrong terms or over-translation can significantly distort the intended meaning of a sentence. These errors indicate a misalignment between the source text and its translation, often leading to inaccurate quality estimation. Mistranslations misrepresent the original text's meaning while missing information indicates an incomplete transfer of content from the source to the target language. Both types of errors lead to inadequately capturing the source text's context in the translation which leads to poor QE predictions. When translations are excessively long or abruptly cut off (missing information), it can complicate the model's ability to perform accurate quality estimation. Lengthy texts may contain more complexity or redundant information, increasing the likelihood of errors. Incomplete sentences fail to provide a full context, making it difficult for QE models to accurately gauge translation quality.

5.1 On DA Distribution

In addition to the initial error analysis, further investigation was conducted for low-resource languages against high-resource languages, inferencing from the best performing model MonoTQ-InfoXLM-large from Experiment 4 - Setting II, detailed in section 3.3.

We choose En-Ta (English-Tamil) in the low-resource category and En-Zh (English-Chinese) in the high-resource category. The distribution between the predicted and true values can be seen in Appendix H. The overall correlation scores for the En-Ta pair are Pearson: 0.727 and Spearman: 0.685, indicating a strong correlation. In contrast, the En-Zh pair shows lower scores, with a Pearson Correlation of 0.475 and a Spearman Correlation of 0.464, suggesting a weaker

Table 4. Comparison of data distribution across buckets for En-Zh and En-Ta.

Bucket Distribution	DA score range	En-Zh Count	En-Ta Count
Bucket 1	0-20	0	0
Bucket 2	21-40	40	4
Bucket 3	41-60	271	52
Bucket 4	61-80	636	278
Bucket 5	81-100	9	630

correlation. The low correlation scores for the high-resource En-Zh language pair were unexpected, challenging typical expectations of language resource impact on model performance.

So we further continued our analysis to find the DA score distribution among En-Ta and En-Zh by dividing our test set into five buckets based on the DA score ranges (0-20, 21-40, 41-60, 61-80, 81-100). Then we calculated the correlation for each bucket for both the language pairs which is shown in figure 3. The bucket-wise correlation scores for the En-Ta pair demonstrated stronger correlations than those for En-Zh, a high-resource language.

We observe the dataset distribution for each bucket as shown in Table 4. It can be noticed that the En-Ta dataset is more skewed towards higher DA score buckets, indicating fewer errors in the translations. Conversely, the En-Zh dataset exhibits lower frequency in the highest bucket. This might be a reason for the lower correlation of En-Zh. This suggests that the test sets for each DA score bucket need to be more equally distributed and for the En-Ta language pair, we need a more challenging, relatively well-distributed test set.

6 Conclusion

This paper is focused on analysing the efficiency of the low-resourced Indic language QE models from the perspective of how language relatedness impacts the efficiency of pre-trained encoder-based QE models.

Our experimental findings corroborate that multilingual proficiency significantly strengthens the performance of QE models, particularly when the languages in question are members of the same linguistic family. Among a comprehensive series of experiments, the MonoTQ-InfoXLM-large model, leveraging a contrastive learning technique, emerged as the most effective model in comparison to other encoder-based models for each language pair except En-Te.

Future directions for our research include expanding our investigation into quality estimation through the utilization of state-of-the-art large language models. This will involve conducting additional experiments that cover a wider array of low-resourced languages. The error analysis conducted reveals a concentration of mistakes within specific categories. Future efforts will focus on a deeper examination of these errors to develop strategies aimed at mitigating

them. Furthermore, our investigation will concentrate on employing efficient model ensemble methodologies for Quality Estimation in low-resource language scenarios.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Baek, Y., Kim, Z.M., Moon, J., Kim, H., Park, E.: Patquest: Papago translation quality estimation. In: Proceedings of the Fifth Conference on Machine Translation. pp. 991–998 (2020)
2. Baek, Y., Kim, Z.M., Moon, J., Kim, H., Park, E.: PATQUEST: Papago translation quality estimation. In: Proceedings of the Fifth Conference on Machine Translation. pp. 991–998. Association for Computational Linguistics, Online (Nov 2020), https://aclanthology.org/2020.wmt-1.113
3. Bao, K., Wan, Y., Liu, D., Yang, B., Lei, W., He, X., Wong, D.F., Xie, J.: Alibaba-translate china's submission for wmt 2022 quality estimation shared task. arXiv preprint arXiv:2210.10049 (2022)
4. Blain, F., Zerva, C., Ribeiro, R., Guerreiro, N.M., Kanojia, D., C. de Souza, J.G., Silva, B., Vaz, T., Jingxuan, Y., Azadi, F., Orasan, C., Martins, A.: Findings of the WMT 2023 shared task on quality estimation. In: Koehn, P., Haddow, B., Kocmi, T., Monz, C. (eds.) Proceedings of the Eighth Conference on Machine Translation. pp. 629–653. Association for Computational Linguistics, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.wmt-1.52, https://aclanthology.org/2023.wmt-1.52
5. Chi, Z., Dong, L., Wei, F., Yang, N., Singhal, S., Wang, W., Song, X., Mao, X.L., Huang, H., Zhou, M.: Infoxlm: An information-theoretic framework for cross-lingual language model pre-training. arXiv preprint arXiv:2007.07834 (2020)
6. Cohen, I., Huang, Y., Chen, J., Benesty, J., Benesty, J., Chen, J., Huang, Y., Cohen, I.: Pearson correlation coefficient. Noise reduction in speech processing pp. 1–4 (2009)
7. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 8440–8451. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.747, https://aclanthology.org/2020.acl-main.747
8. Geng, X., Zhang, Y., Huang, S., Tao, S., Yang, H., Chen, J.: Njunlp's participation for the wmt2022 quality estimation shared task. In: Proceedings of the Seventh Conference on Machine Translation (WMT). pp. 615–620 (2022)
9. Guzmán, F., Chen, P.J., Ott, M., Pino, J., Lample, G., Koehn, P., Chaudhary, V., Ranzato, M.: The flores evaluation datasets for low-resource machine translation: Nepali-english and sinhala-english. arXiv preprint arXiv:1902.01382 (2019)
10. Ive, J., Blain, F., Specia, L.: Deepquest: a framework for neural-based quality estimation. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 3146–3157 (2018)

11. Johnson, M., Schuster, M., Le, Q.V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., Dean, J.: Google's multilingual neural machine translation system: Enabling zero-shot translation. Transactions of the Association for Computational Linguistics **5**, 339–351 (2016). https://doi.org/10.1162/tacl_a_00065
12. Kepler, F., Trénous, J., Treviso, M., Vera, M., Góis, A., Farajian, M.A., Lopes, A.V., Martins, A.F.T.: Unbabel's participation in the WMT19 translation quality estimation shared task. In: Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2). pp. 78–84. Association for Computational Linguistics, Florence, Italy (Aug 2019). https://doi.org/10.18653/v1/W19-5406, https://aclanthology.org/W19-5406
13. Kepler, F., Trénous, J., Treviso, M., Vera, M., Martins, A.F.: Openkiwi: An open source framework for quality estimation. arXiv preprint arXiv:1902.08646 (2019)
14. Kim, H., Jung, H.Y., Kwon, H., Lee, J.H., Na, S.H.: Predictor-estimator: neural quality estimation based on target word prediction for machine translation. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP) **17**(1), 1–22 (2017)
15. Kocmi, T., Federmann, C.: Large language models are state-of-the-art evaluators of translation quality. arXiv preprint arXiv:2302.14520 (2023)
16. Lample, G., Ott, M., Conneau, A., Denoyer, L., Ranzato, M.: Phrase-based neural unsupervised machine translation pp. 5039–5049 (2018). https://doi.org/10.18653/v1/D18-1549
17. Liang, D., Gonen, H., Mao, Y., Hou, R., Goyal, N., Ghazvininejad, M., Zettlemoyer, L., Khabsa, M.: Xlm-v: Overcoming the vocabulary bottleneck in multilingual masked language models. arXiv preprint arXiv:2301.10472 (2023)
18. Lim, S., Park, J.: Papago's submission to the wmt22 quality estimation shared task. In: Proceedings of the Seventh Conference on Machine Translation (WMT). pp. 627–633 (2022)
19. Moura, J., Vera, M., van Stigt, D., Kepler, F., Martins, A.F.: Ist-unbabel participation in the wmt20 quality estimation shared task. In: Proceedings of the Fifth Conference on Machine Translation. pp. 1029–1036 (2020)
20. Perrella, S., Proietti, L., Scirè, A., Campolungo, N., Navigli, R.: Matese: Machine translation evaluation as a sequence tagging problem. In: Proceedings of the Seventh Conference on Machine Translation (WMT). pp. 569–577 (2022)
21. Ranasinghe, T., Orasan, C., Mitkov, R.: Transquest: Translation quality estimation with cross-lingual transformers. arXiv preprint arXiv:2011.01536 (2020)
22. Rei, R., Treviso, M., Guerreiro, N.M., Zerva, C., Farinha, A.C., Maroti, C., C. de Souza, J.G., Glushkova, T., Alves, D., Coheur, L., Lavie, A., Martins, A.F.T.: CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In: Koehn, P., Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussà, M.R., Federmann, C., Fishel, M., Fraser, A., Freitag, M., Graham, Y., Grundkiewicz, R., Guzman, P., Haddow, B., Huck, M., Jimeno Yepes, A., Kocmi, T., Martins, A., Morishita, M., Monz, C., Nagata, M., Nakazawa, T., Negri, M., Névéol, A., Neves, M., Popel, M., Turchi, M., Zampieri, M. (eds.) Proceedings of the Seventh Conference on Machine Translation (WMT). pp. 634–645. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid) (Dec 2022), https://aclanthology.org/2022.wmt-1.60
23. Rei, R., Treviso, M., Guerreiro, N.M., Zerva, C., Farinha, A.C., Maroti, C., de Souza, J.G., Glushkova, T., Alves, D.M., Lavie, A., et al.: Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task. arXiv preprint arXiv:2209.06243 (2022)

24. Sedgwick, P.: Spearman's rank correlation coefficient. Bmj **349** (2014)
25. Sindhujan, A., Kanojia, D., Orasan, C., Ranasinghe, T.: SurreyAI 2023 submission for the quality estimation shared task. In: Koehn, P., Haddow, B., Kocmi, T., Monz, C. (eds.) Proceedings of the Eighth Conference on Machine Translation. pp. 849–855. Association for Computational Linguistics, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.wmt-1.74, https://aclanthology.org/2023.wmt-1.74
26. Specia, L., Blain, F., Fomicheva, M., Zerva, C., Li, Z., Chaudhary, V., Martins, A.F.T.: Findings of the WMT 2021 shared task on quality estimation. In: Proceedings of the Sixth Conference on Machine Translation. pp. 684–725. Association for Computational Linguistics, Online (Nov 2021), https://aclanthology.org/2021.wmt-1.71
27. Specia, L., Paetzold, G.H., Scarton, C.: Multi-level Translation Quality Prediction with QUEST++. pp. 115–120. Association for Computational Linguistics and The Asian Federation of Natural Language Processing (7 2015). https://doi.org/10.3115/v1/P15-4020, https://aclanthology.org/P15-4020
28. Specia, L., Scarton, C., Paetzold, G.H.: Quality estimation for machine translation. Synthesis Lectures on Human Language Technologies **11**(1), 1–162 (2018)
29. Specia, L., Shah, K., De Souza, J.G., Cohn, T.: Quest-a translation quality estimation framework. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 79–84 (2013)
30. Tao, S., Chang, S., Miaomiao, M., Yang, H., Geng, X., Huang, S., Zhang, M., Guo, J., Wang, M., Li, Y.: CrossQE: HW-TSC 2022 submission for the quality estimation shared task. In: Koehn, P., Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussà, M.R., Federmann, C., Fishel, M., Fraser, A., Freitag, M., Graham, Y., Grundkiewicz, R., Guzman, P., Haddow, B., Huck, M., Jimeno Yepes, A., Kocmi, T., Martins, A., Morishita, M., Monz, C., Nagata, M., Nakazawa, T., Negri, M., Névéol, A., Neves, M., Popel, M., Turchi, M., Zampieri, M. (eds.) Proceedings of the Seventh Conference on Machine Translation (WMT). pp. 646–652. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid) (Dec 2022), https://aclanthology.org/2022.wmt-1.61
31. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
32. Wenzek, G., Lachaux, M.A., Conneau, A., Chaudhary, V., Guzmán, F., Joulin, A., Grave, E.: CCNet: Extracting high quality monolingual datasets from web crawl data. In: Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Twelfth Language Resources and Evaluation Conference. pp. 4003–4012. European Language Resources Association, Marseille, France (May 2020), https://aclanthology.org/2020.lrec-1.494
33. Zerva, C., Blain, F., Rei, R., Lertvittayakumjorn, P., C. de Souza, J.G., Eger, S., Kanojia, D., Alves, D., Orăsan, C., Fomicheva, M., Martins, A.F.T., Specia, L.: Findings of the WMT 2022 shared task on quality estimation. In: Proceedings of the Seventh Conference on Machine Translation (WMT). pp. 69–99. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid) (Dec 2022), https://aclanthology.org/2022.wmt-1.3

A Appendix: Direct assessment score dataset

Table 5. DA score dataset distribution across different language pairs and test sets. Train, Dev and Test columns show the number of datasets distributed for training, evaluation and testing. The primary experiments were conducted using the language pairs English-Marathi (En-Mr), English-Gujarati (En-Gu), English-Hindi (En-Hi), English-Tamil (En-Ta), and English-Telugu (En-Te)

	DA Score		
Lang.	Train	Dev	Test
English - Marathi (En-Mr)	27 000	1000	1086
English - Gujarati (En-Gu)	7000	1000	1075
English - Hindi (En-Hi)	7000	1000	1074
English - Tamil (En-Ta)	7000	1000	1067
English - Telugu (En-Te)	7000	1028	1000
English - German (En-De)	7000	1000	1000
English - Chinese (En-Zh)	7000	1000	1000
Russian - English (Ru-En)	7000	1000	1000
Romanian - English (Ro-En)	7000	1000	1000
Nepalis - English (Ne-En)	7000	1000	1000
Sinhala - English (Si-En)	7000	1000	1000

B Appendix: Diskfootprint comparison

Table 6. Row I and II shows the disk footprint of the best-performing systems of WMT23 QE shared task for Indic languages. Meanwhile, rows III to V display the disk footprint of our models.

Row	**Name**	**Disk Foot Print (Bytes)**
I	Unbabel-IST	42,868,104,221
II	HW-TSC	27,730,527,504
III	MonoTQ-XLMV	3,221,225,472
IV	MonoTQ-InfoXLM-large	2,362,232,012
V	MonoTQ-XLMR-large	2,254,857,830

C Appendix: Scatter plots depicting the distribution of predicted quality score and true value

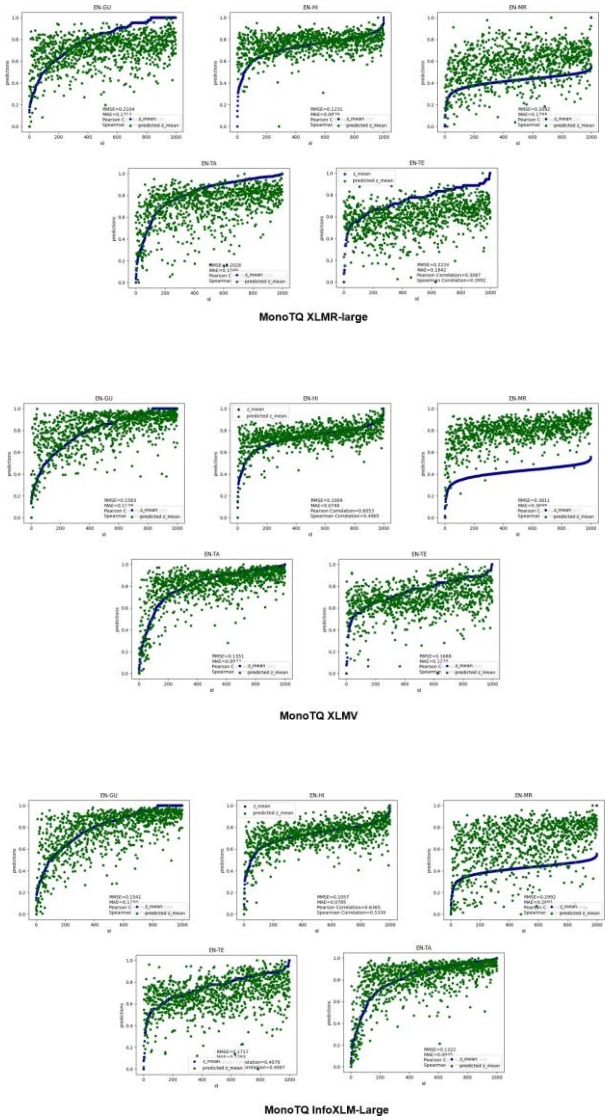

Fig. 4. Scatter plots depicting the distribution of predicted quality score versus true values of Experiment 1 with development dataset for Indic-language-only multilingual setting using the MonoTQ-XLMR-large, MonoTQ-XLMV and MonoTQ-InfoXLM-large models

D Experiment results of full, few and zero-shot settings

Table 7. Spearman (ρ) and Pearson (r) correlation scores for different model configurations and settings where Indic language datasets for DA scores are used for training and testing. The highest performance score obtained for each language pair in each setting is marked in bold

Model	En-Gu		En-Hi		En-Mr		En-Ta		En-Te	
	r	ρ	*r*	ρ	*r*	ρ	*r*	ρ	*r*	ρ
I. Zero-Shot										
MonoTQ-XLMR-large	0.187	0.209	0.222	0.252	**0.313**	0.308	0.173	0.291	**0.193**	**0.241**
MonoTQ-InfoXLM-large	**0.231**	**0.239**	**0.266**	**0.288**	0.308	**0.331**	**0.387**	**0.487**	0.189	0.226
MonoTQ-XLMV	0.150	0.188	0.186	0.188	0.292	0.241	0.252	0.337	0.060	0.090
II. Few-Shot - 50 training samples										
MonoTQ-XLMR-large	**0.480**	**0.613**	**0.425**	**0.451**	0.096	0.109	0.104	0.182	-0.077	-0.041
MonoTQ-InfoXLM-large	0.435	0.423	0.375	0.350	0.249	**0.294**	0.021	0.103	-0.081	-0.047
MonoTQ-XLMV	-0.015	0.045	0.041	0.083	**0.289**	0.212	**0.272**	**0.203**	**0.085**	**-0.142**
III. Few-Shot - 100 training samples										
MonoTQ-XLMR-large	0.401	0.398	0.435	0.459	0.181	0.210	0.066	0.107	-0.008	-0.037
MonoTQ-InfoXLM-large	**0.437**	**0.474**	**0.500**	0.501	**0.207**	**0.269**	0.014	0.073	**0.174**	**0.167**
MonoTQ-XLMV	0.426	0.411	0.481	**0.503**	-0.135	-0.126	**0.110**	**0.150**	0.043	0.079
IV. Few-Shot - 200 training samples										
MonoTQ-XLMR-large	**0.383**	**0.394**	**0.448**	**0.522**	**0.050**	**0.087**	**0.161**	**0.218**	**0.055**	**0.051**
MonoTQ-InfoXLM-large	0.307	0.323	0.358	0.356	0.000	0.005	0.159	0.183	0.042	-0.075
MonoTQ-XLMV	-0.119	-0.097	0.019	0.089	-0.040	0.011	-0.180	-0.138	-0.033	-0.057
V. Full training Data										
MonoTQ-XLMR-large	0.300	0.438	0.430	0.440	-0.117	0.395	0.454	0.482	0.211	0.345
MonoTQ-InfoXLM-large	**0.656**	**0.713**	**0.726**	**0.624**	0.030	0.470	**0.662**	**0.726**	**0.719**	0.462
MonoTQ-XLMV	0.536	0.673	0.687	0.572	**0.426**	**0.642**	0.559	0.670	0.642	**0.464**

This experiment analyses sample efficiency by investigating the performance across zero-shot, few-shot, and full-data scenarios. MonoTQ-InfoXLM-large generally exhibits the most robust performance across the languages in the zero-shot setting (I). MonoTQ-XLMR-large is competitive and, in some cases, very close to the leading model. The MonoTQ-XLMR-large model appears to gain more consistent advantages from increased sample sizes in few-shot setting (II, III, IV), with this trend being particularly evident in all the language pairs. The findings from few-shot experiments suggest that the efficacy of few-shot learning in quality estimation is highly contingent on both the model selection and the sample size. In the full data setting (V) which has all the indic language pairs in the training dataset, MonoTQ-InfoXLM-large is consistently the top-performing model for almost all language pairs except En-Mr. Performance degradation is noticeable from full data to few-shot to zero-shot settings, which is expected given the decrease in available training data. Generally, the MonoTQ-InfoXLM-large model shows robust performance across all the different settings and language pairs, which highlights the contrastive learning technique shows better sample efficiency.

E Appendix: Heatmaps for the correlation scores from experiment 1 & 4

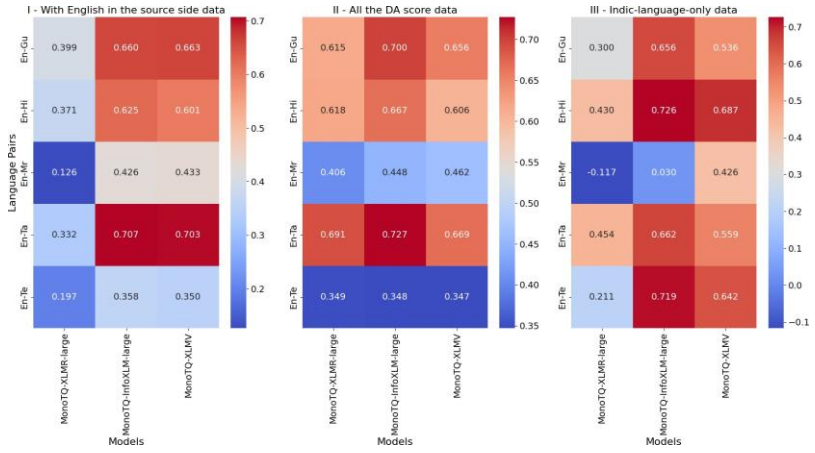

Fig. 5. Comparative heatmap of the Pearson correlation scores for different models of Experiments 4 (I - with English in the source side data combined with Indic language for training, II - All the DA score data combined for training) compared with Experiment 1 (III - Indic-language-only-multilingual setting)

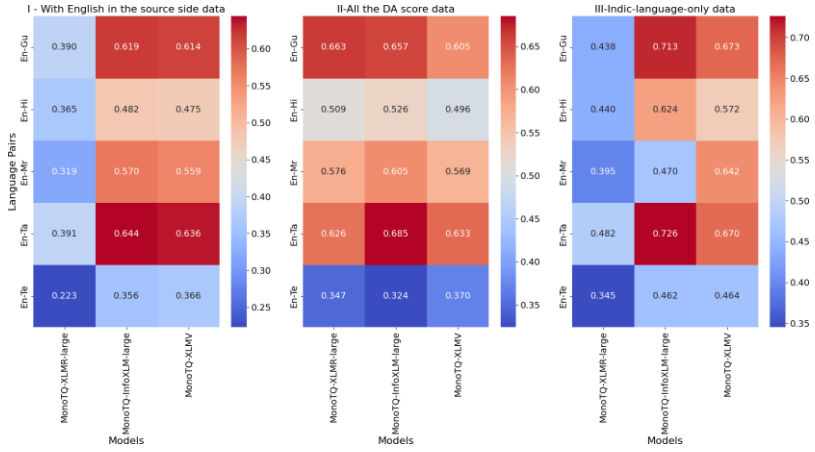

Fig. 6. Comparative heatmap of the Spearman correlation scores for different models of Experiments 4 (I - with English in the source side data combined with Indic language for training, II - All the DA score data combined for training) compared with Experiment 1 (III - Indic-language-only-multilingual setting)

F Appendix: Overall analysis of the experiments

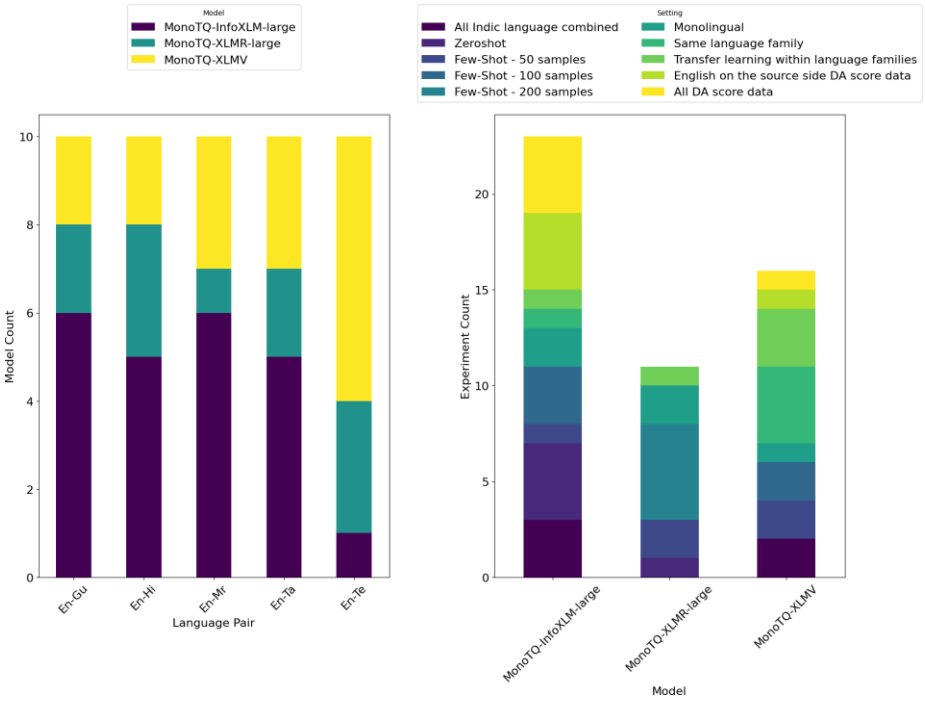

Fig. 7. The left image displays the frequency of the highest-performing model by language pair across all experiments, while the right image illustrates the count of instances where each model showed the best performance in each experimental setting.

G Appendix: Translation Error Categories and Examples

Table 8. The following table shows some examples of translation error types which led to the highest deviations between the true and predicted quality scores explained in section 5. S -> Source, T -> Translation, Tr -> Transliteration, ET-> English translation of the machine-translated Tamil sentence (T).

Use of Entity
S - Participation in International Quality Assurance Programme conducted by WHO Collaborating Centre functioning at Australia. **T** - ஆஸ்தி◌ேரலியாவ¥ல் ெசயல்படும் உலக சுகாதார அைமப்ப¥ன் ஒத்துைழப்பு ைமயத்தால் நடத்தப்படும் சர்வேதச தர உறுதி◌த் தி◌ட்டத்தி◌ல் பங்◌ேகற்பு. **Tr**- Āstirēliyāvil (in Australia) ceyalpaṭum (functioning) ulaka (world) cukātāra (health) amaippiṉ (of the organization) ottuḻaippu (collaboration) maiyattāl (by the center) naṭattappaṭum (conducted) carvatēca (international) tara (quality) uṟutit (assurance) tiṭṭattil (in the program) paṅkēṟpu (participation). **ET** - Participation in an International Quality Assurance Programme conducted by the World Health Organization Collaborating Center in Australia.
Transliteration
S - The cable car station is in the middle of the wall close to tower 14. **T** - ◌ேகப¥ள் கார் நி◌ைலயம் ◌ேகாபுரம் 14 க்கு அருகி◌ல் சுவரின் நடுவ¥ல் உள்ளது. **Tr**- Kēpiḷ kār (cable car) nilaiyam (station) kōpuram (tower) 14-kku (to 14) arukil (close) cuvarin (of the wall) naṭuvil (in the middle) uḷḷatu (is located) **ET** - The cable car station is in the middle of the wall close to tower 14.
Wrong terms
S - I think it's better if we encourage our great creative minds to live. **T** - நமது பைடப்பாற்றல் மிக்க மனங்கைள வாழச் ெசய்தால் நல்லது என்று நான் நி◌ைனக்கி◌ேறன். **Tr**- Namatu (our) paṭaippāṭṭal (creativity) mikka (great) manaṅkaḷai (minds) vāḻac (to live) ceytāl (if done) nallatu (good) eṉṟu (as) nāṉ (I) niṉaikkiṟēṉ (think) **ET** - I think it's good to let our creative minds live.
Overtranslation
S - I was so worried. **T** - இதனால் நான் மிகவும் கவைலயைடந்◌ேதன். **Tr**- Ithanal (because of this) nān (I) mikavum (very much) kavalaiyadainthēn (was worried) **ET** - I was so worried because of this
Mistranslation
S - By the end of their study, they were satisfied that they had an authentic tsantsa. **T** - அவர்களுைடய படிப்ப¥ன் முடிவ¥ல், தங்களுக்கு உண்ைமயான சான்ட்ஸா இருப்பைத அவர்கள் தி◌ருப்தி◌ ெசய்தனர். **Tr**- Avarkaḷuṭaiya (their) paṭippiṉ (of the study) muṭivil (at the end), taṅkaḷukku (to them) uṇmaiyāṉa (real) cāṉṭsā (tsantsa) iruppatai (having) avarkaḷ (they) tirupti (satisfaction) ceytaṉar (achieved). **ET** - At the end of their study, they satisfied they had the real tsantasa.
Missing Information
S - Bulman and Camille interviewed 29 people suffering from paralysis caused by different types of automobile accidents, or a fall or by injury in the football field. **T** - பல்◌ேமனும் கமிலும் 29 ◌ேபைர ◌ேபட்டி கண்டனர். **Tr**- Palmenum (Bulman and) Camilum (Camille) 29 (twenty-nine) perai (people) petti (interviewed) kantanar (they saw/discussed) **ET** - Bulman and Camille interviewed 29 people

H Appendix: Error analysis scatter-plot graphs

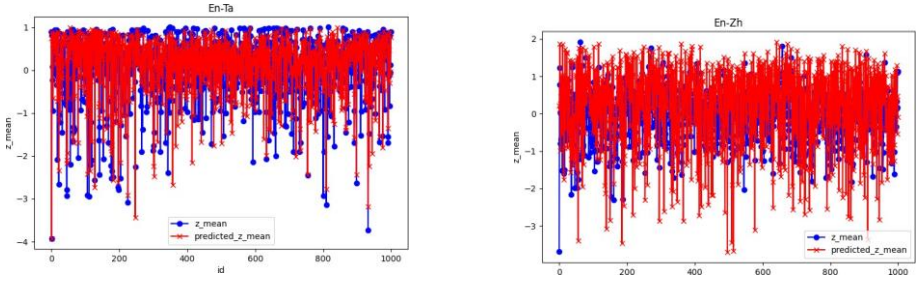

Fig. 8. Scatter plots depicting the distribution of predicted quality score vs true values of high-resource (en-zh) and low-resource (en-ta) languages