When LLMs Struggle: Reference-less Translation Evaluation for Low-resource Languages

Archchana Sindhujan, **Diptesh Kanojia**, **Constantin Orăsan**, **Shenbin Qian**

Institute for People-Centred AI and Centre for Translation Studies,
University of Surrey, United Kingdom
{a.sindhujan, d.kanojia, c.orasan, s.qian}@surrey.ac.uk

Abstract

This paper investigates the *reference-less* evaluation of machine translation for low-resource language pairs, known as quality estimation (QE). Segment-level QE is a challenging cross-lingual language understanding task that provides a quality score ($0 - 100$) to the translated output. We comprehensively evaluate large language models (LLMs) in zero/few-shot scenarios and perform instruction fine-tuning using a novel prompt based on annotation guidelines. Our results indicate that prompt-based approaches are outperformed by the encoder-based fine-tuned QE models. Our error analysis reveals tokenization issues, along with errors due to transliteration and named entities, and argues for refinement in LLM *pre-training* for cross-lingual tasks. We release the data, and models trained publicly for further research.

1 Introduction

Traditional methods of obtaining references for machine-translated texts are costly, and prone to subjectivity and inconsistency (Rei et al., 2021; Lo et al., 2014; Huynh et al., 2008). To address these challenges of evaluating imperfect translations, Quality Estimation (QE) has emerged as a crucial area, enabling the assessment of MT output in the absence of a reference (Zerva et al., 2022).

Our work investigates segment-level QE (Blain et al., 2023; Zerva et al., 2022; Fernandes et al., 2023), which is *conventionally* modelled as a *regression task* and aims to predict a segment-level quality score, also known as the direct assessment (DA) score (Graham et al., 2013). Due to the underlying subjectivity in human translation quality evaluation, DA score is computed as a *mean* of three or more human annotations on a scale of $0 - 100$. While large language models (LLMs) claim superlative performance for different natural language processing (NLP) tasks (Devlin et al., 2019; Achiam et al., 2023), evaluation of machine-translated output poses a unique challenge where both *syntactic accuracy and cross-lingual semantic match* are relevant, for the prediction of DA scores.

LLMs are applicable for many NLP tasks, including machine translation (MT) (Kocmi et al., 2023; Robinson et al., 2023; Manakhimova et al., 2023) and quality estimation (Kocmi and Federmann, 2023; Xu et al., 2023; Fernandes et al., 2023; Huang et al., 2024). There are significant disparities in the reported performance of LLMs between high- and low-resource languages (Huang et al., 2023; Nguyen et al., 2024). LLMs exhibit better performance in evaluating the quality when references are available (Huang et al., 2024); however, they are challenging to scale due to the cost associated with manual translation.

This work focuses on the *reference-less* scenario, evaluating the efficacy of LLMs in settings like zero-shot, few-shot/in-context learning (ICL), and instruction fine-tuning with an adapter (Hu et al., 2021). We present a novel prompt which utilizes annotation guidelines within prompt instructions and improves task performance. Additionally, we perform experiments for both independent language-pair training (*ILT - training instances from one language pair*), and unified multilingual training (*UMT - training instances from all language pairs*) settings. Our contributions are:

- A novel annotation guidelines-based prompt (AG-prompt) which improves zero-shot performance.
- A comprehensive evaluation for segment-level QE using multiple LLMs, indicating challenges for cross-lingual NLP tasks.
- Instruction fine-tuned QE model adapters *(4-bit)* for quick deployment.
- Quantitative and Qualitative analysis indicating critical challenges using LLMs for cross-lingual tasks involving low-resource languages.

Proceedings of the First Workshop on Language Models for Low-Resource Languages (LoResLM 2025), pages 473–495
January 20, 2025.

2 Background

Transformer-based approaches which leverage supervised fine-tuning of regression models significantly improved the performance of QE models (Ranasinghe et al., 2020). Recently proposed approaches like CometKiwi (Rei et al., 2023), Ensemble-CrossQE (Li et al., 2023) and TransQuest (Ranasinghe et al., 2020; Sindhujan et al., 2023) from WMT QE shared tasks (Blain et al., 2023) are based on pre-trained encoder-based language models. However, recent claims have propelled the use of LLMs across various NLP tasks (Zhao et al., 2023). Following suit, Kocmi and Federmann (2023) introduced the GEMBA prompt-based metric for evaluating translation quality. Their approach focuses on zero-shot prompt-based evaluation, comparing four prompt variants across nine GPT model variants for three high-resource language pairs. The paper discusses experiments with both settings, with and without reference, claiming SoTA performance by including the reference for DA prediction. Our experiments reproduce their prompt in a reference-less setting utilizing only publicly available LLMs and compare prompting strategies by adding relevant context.

Huang et al. (2024) examined how LLMs use source and reference information for translation evaluation and they observed that reference information improves accuracy and correlations, while source information shows a negative impact, highlighting limitations in LLMs' cross-lingual semantic matching capability, which is essential for a task such as QE. Mujadia et al. (2023) perform QE by pre-tuning the adapter using a large parallel corpus of English-Indic languages over machine translation task. They fine-tune the model again using supervised QE data and show that pre-tuning the model using MT does not help. Other approaches to QE such as MQM, include fine-grained error annotation and detailed explanations, which are often not viable for low-resource languages due to lack of annotated data.

3 Methodology

3.1 Datasets

Our study focuses on low-resource language pairs from the WMT QE shared tasks with human-annotated DA scores, including English to Gujarati, Hindi, Marathi, Tamil, and Telugu (En-Gu, En-Hi, En-Mr, En-Ta, En-Te) from WMT23 (Blain et al., 2023). We also include Estonian, Nepali, and Sinhala to English (Et-En, Ne-En, Si-En) language pairs from WMT22 (Zerva et al., 2022). Hindi and Estonian although mid-resource for machine translation (Nguyen et al., 2024), lack sufficient resources for translation evaluation and QE. Training splits were used for fine-tuning, while test splits were used for zero-shot, ICL, and inference experiments (Appendix C).

3.2 Prompting Strategies

Zero-shot prompting refers to a model generating outputs for a given input prompt solely based on its pre-trained knowledge and inherent generalization capabilities, without requiring any additional fine-tuning or contextual examples. Existing studies highlight that adding context and reasoning to prompts can significantly enhance LLM's performance in NLP tasks (Zhou et al., 2023; Chen et al., 2023). However, for the low-resource languages, fine-grained QE data is unavailable. Therefore, we experiment with different prompting strategies: 1) instructing the model to act as a translation evaluator (TE) (Appendix B) and 2) providing additional context from human annotation guidelines (AG). Using the proposed AG prompt (Figure 1), we incorporate reasoning to evaluate translation quality. We compare these strategies with the GEMBA prompt (Kocmi and Federmann, 2023) in the zero-shot setting.

In-context learning refers to the ability of large language models to perform a task by leveraging examples of the task provided within the input context, without requiring any additional training. We focus our investigation on the AG prompt within the ICL scenario. In this setting, we augmented the AG prompt with example annotations from 5 different DA score ranges (0-30, 31-50, 51-70, 71-90, 91-100), as detailed in Appendix A. The ICL experiments were divided into three configurations: 3-ICL, 5-ICL, and 7-ICL. In the 5-ICL configuration, we selected one example from each of the five predefined DA score ranges. The 3-ICL configuration excluded examples from the 31-50 and 51-70 ranges. For the 7-ICL configuration, we included one example from each range, plus two additional samples—one from the lowest and one from the highest available score ranges. Through in-context learning experiments, we aim to assess whether

We need to Evaluate the machine translated sentences of **<Source language>**(Source) to **<Target language>** (Translation), with quality scores ranging from 0 to 100.

Source: <Source Sentence>

Translation: <Translated Sentence>

Scores of 0-30 indicate that the translation is mostly unintelligible, either completely inaccurate or containing only some keywords. Scores of 31-50 suggest partial intelligibility, with some keywords present but numerous grammatical errors. A score between 51-70 means the translation is generally clear, with most keywords included and only minor grammatical errors. Scores of 71-90 indicate the translation is clear and intelligible, with all keywords present and only minor non-grammatical issues. Finally, scores of 91-100 reflect a perfect or near-perfect translation, accurately conveying the source meaning without errors.

The evaluation criteria focus on two main aspects: Adequacy (how much information is conveyed) and Fluency (how grammatically correct the translation is). Predict the quality score in the range of 0 to 100 considering the above instructions. Predict only the score, no need for explanation.

Score:

Figure 1: The proposed AG prompt which augments scoring instructions within the context.

incorporating examples of DA annotations can enhance the model's performance. Additionally, by varying the number of examples in each ICL setting, we investigate the impact on the performance of the QE model.

Furthermore, instruction fine-tuning involves adapting a model using a dataset that includes explicit instructions for specific tasks. In our instruction fine-tuning experiments, we employ the AG prompt to evaluate its effect on model performance.

3.3 Implementation Details

For our study, we focus on publicly available LLMs with a parameter count under 13B that have established benchmarks in multilingual performance: Gemma-7B[1], OpenChat-3.5[2], Llama-2-7B[3], Llama-2-13B[4]

The OpenChat 7B-parameter model (Wang et al., 2023) (OC-3.5-7B) employs Conditioned-RLFT, a technique that uses a class-conditioned policy to prioritize high-quality responses over sub-optimal ones. The Llama model (Touvron et al., 2023) incorporates supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF) to align its outputs with human preferences. Additionally, the Gemma-7B model (Mesnard et al., 2024) utilizes advanced techniques such as Multi-Query Attention, RoPE Embeddings, GeGLU Activations, and RMSNorm to enhance its performance. We chose not to use the latest Llama models (Llama-3 and Llama-3.1) in our experiments, as results from initial zero-shot evaluations showed they did not produce meaningful outputs.

We fine-tune regression models using QE frameworks such as TransQuest (Ranasinghe et al., 2020), in both Independent Language-Pair Training and Unified Multilingual Training settings. For comparison, we use the COMET model (Rei et al., 2023), which is fine-tuned on low-resource language pairs (mentioned in the section 3.1) utilizing the pre-trained encoder transformer XLM-R-XL (Goyal et al., 2021). We chose to restrict the investigation to zero-shot, in-context learning and adapter fine-tuning. Approaches which use continual pre-training are not within the scope of this investigation due to their computational cost, leaving them for future work.

Zero-shot and ICL scenarios We utilize the vLLM framework (Kwon et al., 2023) to perform our experiments. For all our zero-shot and ICL experiments, we experimented with the default temperature value of 0.85 and also the value of 0. The temperature value of 0 provided a more stable and consistent output. The input sequence length was set to 1024 for zero-shot inference and 4096 for ICL inference.

Instruction fine-tuning We used the LLaMA-Factory framework (Zheng et al., 2024) for fine-tuning experiments, leveraging its prompt formatting capabilities. For efficient tuning, we applied LoRA (Hu et al., 2021), focusing on the query and value projection layers of transformers,

[1] huggingface.co/google/Gemma-7b
[2] huggingface.co/OpenChat/OpenChat-3.5
[3] huggingface.co/meta-llama/LLaMA-2-7b-chat-hf
[4] huggingface.co/meta-llama/LLaMA-2-13b-chat-hf

LP	Template	Gemma-7B	Llama-2-7B	Llama-2-13B	OC-3.5-7B
En-Gu	0-shot-GEMBA	0.113	0.006	0.019	0.249*
	0-shot-TE	-0.102†	-0.008	-0.052	0.117†
	0-shot-AG	-0.079	-0.007	0.008	0.164†
	3-ICL-AG	-0.005	0.036	-0.036	0.223
	5-ICL-AG	0.023	-0.008	0.095	0.151
	7-ICL-AG	0.071	-0.053	-0.108	**0.260**
En-Hi	0-shot-GEMBA	0.131	-0.002	0.009	**0.254***
	0-shot-TE	-0.050	-0.072	0.056	0.134
	0-shot-AG	-0.056	-0.029	0.069	0.253
	3-ICL-AG	0.134	-0.114	-0.023	0.184
	5-ICL-AG	0.075	-0.022	0.035	0.212
	7-ICL-AG	0.075	-0.176	0.014	0.163
En-Mr	0-shot-GEMBA	0.135	0.053	0.115	0.183
	0-shot-TE	0.173	0.070	0.040	0.114
	0-shot-AG	0.027	0.059	0.005	**0.276***
	3-ICL-AG	0.202	0.120	0.095	0.218
	5-ICL-AG	0.164†	0.032	-0.031	0.226
	7-ICL-AG	0.167	0.050	0.047	0.251
En-Ta	0-shot-GEMBA	0.222	0.067	0.091	0.358
	0-shot-TE	-0.037†	0.012	0.016	0.178
	0-shot-AG	-0.002	0.055	-0.070	**0.363***
	3-ICL-AG	0.122	-0.019	0.083	0.337
	5-ICL-AG	0.114	0.017	0.193	0.332
	7-ICL-AG	0.122	-0.096	-0.004	0.309
En-Te	0-shot-GEMBA	0.081	-0.016	0.121†	0.145*
	0-shot-TE	0.018	0.013	0.010	0.072
	0-shot-AG	0.065	0.083	0.045	0.121†
	3-ICL-AG	0.092	0.027	0.015	0.152
	5-ICL-AG	0.021	0.051	0.073	0.126
	7-ICL-AG	-0.033	0.021	-0.028	**0.196**
Et-En	0-shot-GEMBA	0.289	0.168	0.185	0.571
	0-shot-TE	0.086	0.100	0.146	0.455
	0-shot-AG	0.098	0.064	0.319	0.619*
	3-ICL-AG	0.226	0.268	-0.058	0.613
	5-ICL-AG	0.327	0.269	0.438	**0.636**
	7-ICL-AG	0.306	0.033	0.169	0.616
Ne-En	0-shot-GEMBA	0.261	0.153	0.222	0.448
	0-shot-TE	0.155	0.100	0.080	0.334
	0-shot-AG	0.130	0.144	0.303	0.487*
	3-ICL-AG	0.273	0.149	0.340	0.457
	5-ICL-AG	0.305	0.189	0.319	0.471
	7-ICL-AG	0.365	-0.040	0.259	**0.491**
Si-En	0-shot-GEMBA	0.193	0.144	0.195	0.417
	0-shot-TE	0.055	0.129	0.109	0.303
	0-shot-AG	0.042	0.069	0.238	0.441*
	3-ICL-AG	0.306	0.146	0.018	0.470
	5-ICL-AG	0.320†	0.243	0.326	**0.479**
	7-ICL-AG	0.283	-0.017	0.223	0.477

Table 1: Spearman correlation (ρ) between the predicted and human-annotated scores for all the experimental settings. Prompt templates: GEMBA, TE, and AG (from section 3.2). Bold indicates the overall top score per language pair, asterisks (*) denote top scores in zero-shot settings, and underlined values highlight the best among ICL settings. The (†) symbol denotes statistically insignificant results ($p > 0.05$), and the dashed line separates language pairs with English as target.

which proved the most effective in reducing computational cost and memory usage. This approach consistently provided reliable outputs, making these layers our choice for fine-tuning throughout the experiments. We set the LoRA rank to 64, as higher ranks improve adaptation but increase resource demands. To reduce memory usage and speed up inference, we applied 4-bit quantization, with a slight trade-off in accuracy (Dettmers et al., 2023), and used 16-bit floating-point precision (fp16) to enable larger models and batch sizes within the same memory

limits (Micikevicius et al., 2018).

We conducted fine-tuning experiments in two settings: **Unified Multilingual Training (UMT)**, we combined training data from 8 low-resource language pairs (En→Gu, Hi, Mr, Ta, Te and Et, Ne, Si→En) and performed inference using language-specific test sets; **Independent Language-Pair Training (ILT)**, we fine-tuned separate models for each language pair, using individual training data and performing inference with corresponding test sets to evaluate the results. All the AG prompt data[5] used for Instruction Fine-Tuning and evaluation, along with the fine-tuned models, have been publicly released on the HuggingFace platform (Appendix M).

3.4 Evaluation & Metrics

We primarily use Spearman's correlation (Sedgwick, 2014) between the DA mean (averaged human-annotated DA scores from three annotators) and predictions as our evaluation metric. Additionally, Pearson's correlation (Cohen et al., 2009) and Kendall's Tau correlation (Lapata, 2006) are calculated (see Appendices: F, G, I, H).

The predicted outputs from our models contained extra text alongside the predicted DA score, which we extracted using regular expressions. In the zero-shot and ICL experiments, some outputs lacked a score, and those cases were excluded from the correlation analysis (see Appendices F & G). However, this problem is mitigated after instruction fine-tuning where all inferenced instances predicted a score.

Statistical Significance We performed a two-tailed paired T-test to assess statistical significance between human-annotated and predicted DA scores, using a significance threshold of $p < 0.05$. Statistically insignificant results are marked with † in Tables 1 and 2; most other results showed high significance, with $p < 0.01$ or $p < 0.001$.

4 Results

Table 1 presents results from the zero-shot and ICL scenarios. Our proposed AG prompt achieved the highest scores in the zero-shot setting for most language pairs, with the exception of En to{Gu, Hi, Te}. For En to {Hi, Ta} the AG prompt scores were very close to those of the best scores, indicating the AG prompt's strength across the majority of language pairs. Notably, the OpenChat model attained the highest correlation scores for all language pairs in the zero-shot experiment.

Given the AG prompt's strong zero-shot performance, our ICL investigations focused solely on it. In the ICL setting, 4 language pairs (En-Gu, En-Mr, En-Te, Ne-En) performed best with 7-ICL, 3 language pairs (En-Hi, En-Et, Si-En) with 5-ICL, and 1 language pair (En-Ta) with 3-ICL. OpenChat consistently achieved the highest correlation scores across all low-resource pairs, with Et-En, Ne-En, and Si-En outperforming other English-Indic pairs in both zero-shot and ICL.

In Appendix Tables 4, 5, and 6, for the zero-shot setting, we note that the number of dropped rows for the TE prompt is the highest whereas the same when using AG prompts is the lowest, likely because AG prompt specifies the score ranges explicitly.

UMT Setting As shown in Table 2, the OpenChat model achieved the highest correlation scores for En to {Hi, Ta, Te, Si} while Gemma obtained the highest correlation scores for En-{Gu,Mr} and {Et, Ne}-En. However, compared to instruction fine-tuned LLMs, the fine-tuned encoder-based models (TransQuest, CometKiwi) consistently achieved significantly higher correlations among all low-resource language pairs.

ILT Setting As shown in Table 2, OpenChat obtained the best Spearman scores among other LLMs for all the language pairs except En-Mr. Unlike UMT fine-tuning where only pre-trained encoders gave the best result, ILT fine-tuned LLMs achieve the highest results for En to {Hi, Ta, Te} in this setting, where Tamil and Telugu languages are from the Dravidian family which are considered extremely low-resource in terms of pre-training data distribution for LLMs.

Comparing ILT and UMT setting results, the UMT performs better for most low-resource language pairs. This suggests that incorporating diverse linguistic data enhances the model's ability to generalize and accurately evaluate translations across various low-resource languages. Considering the overall best results, fine-tuned encoder-based models demonstrate the best performance.

[5] huggingface.co/datasets/ArchSid/QE-DA-datasets/

Lang-pair	Gemma-7B	Llama-2-7B	Llama-2-13B	OC-3.5-7B	TransQuest	CometKiwi
Unified Multilingual Training (UMT) Setting						
En-Gu	0.566	0.461	0.465	0.554	0.630	**0.637**
En-Hi	0.449	0.332	0.322	0.458	0.478	**0.615**
En-Mr	0.551†	0.516†	0.505	0.545†	**0.606**	0.546
En-Ta	0.502	0.464	0.471	0.509	0.603	**0.635**
En-Te	0.242	0.258	0.258	0.267	**0.358**	0.338
Et-En	0.728	0.636	0.655	0.678	0.760	**0.860**
Ne-En	0.650	0.519	0.565	0.607	0.718	**0.789**
Si-En	0.455	0.395	0.403†	0.481†	0.579	**0.703**
Independent Language-Pair Training (ILT) Setting						
En-Gu	0.440	0.214	0.421	0.520	**0.653**	-
En-Hi	0.375	0.282	0.336	**0.474**	0.119	-
En-Mr	0.557	0.509†	0.501	0.554†	**0.629**	-
En-Ta	0.475	0.375	0.441	**0.509**	0.303	-
En-Te	0.217	0.263	0.261	**0.271**	0.087	-
Et-En	0.648	0.589	0.598	0.652	**0.806**	-
Ne-En	0.612	0.497	0.543†	0.614	**0.746**	-
Si-En	0.387	0.332	0.346	0.441	**0.581**	-

Table 2: Spearman correlation (ρ) scores between the predicted and mean DA scores for *UMT* and *ILT* fine-tuning. For both settings exclusively, scores underlined represent best amongst LLMs, and scores in boldface indicate overall best scores amongst both LLMs and encoder-based models. (†) denotes the statistically insignificant results ($p > 0.05$). The dashed line separates language pairs with English as the target.

5 Discussion

Zero-shot- In comparison to the GEMBA and TE prompts, the AG prompt demonstrated the best overall performance in zero-shot experiments with LLMs across the majority of language pairs. This indicates that in the absence of training data, the additional context provided in the AG prompt—acting as an annotation guide, enhances the effectiveness of LLM-based quality estimation more effectively than LLMs functioning as translation evaluators (TE template) or simply assigning scores based on a straightforward request like in the GEMBA prompt. The structured guidelines in the AG prompt offer a clearer framework for evaluating translation quality, which supports more accurate scoring in zero-shot settings.

ICL- Outperformed zero-shot for most language pairs (En-Gu, En-Te, Et-En, Ne-En, Si-En), suggesting that adding examples improves LLMs' ability to predict translation quality. However, the effect of increasing examples varied across language pairs and models (see Appendix K). When the number of examples in the ICL prompts was increased, the En-Gu and Ne-En language pairs with the Gemma-7B model, as well as the En-Mr and Ne-En language pairs with the OpenChat model, consistently showed improved performance. However, for other language pairs and models, the performance gains were not always evident, suggesting that increasing the number of examples does not necessarily lead to better results.

Fine-tune - We observed a notable improvement in correlation scores when moving from zero-shot to fine-tuning, compared to zero-shot to ICL (Appendix K). This indicates that instruction fine-tuning with task-specific data is more effective than providing detailed examples in prompts. In fine-tuning experiments, pre-trained encoder-based models with UMT settings outperformed LLMs. Despite this, LLMs are significantly larger in size and contain more parameters compared to pre-trained encoder models (Appendix L). While LLMs can handle various NLP tasks and show decent performance in translation evaluation for some low-resource language pairs, they are not specifically trained for regression tasks like pre-trained encoders. This difference likely contributes to LLMs' lower performance in QE. Notably, the OpenChat model consistently outperformed other LLMs when provided with sufficient context as annotation guidelines.

A noteworthy observation is that English, when used as the target language in machine translation, consistently achieved higher correlation scores for QE in zero-shot and ICL experiments with LLMs. Similarly, Figure 2, which highlights *setting-agnostic* best performance for fine-tuned LLMs *vs.* TransQuest-InfoXLM *vs.* COMET,

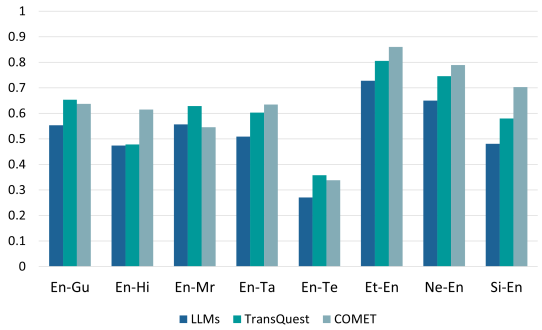

Figure 2: Best fine-tuned performance (Spearman) for LLMs *vs.* TransQuest-InfoXLM *vs.* COMET

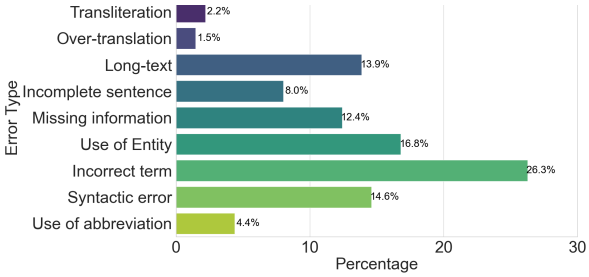

Figure 3: Error types and their percent contribution.

shows enhanced performance with English as the target language and the data distribution for other language pairs is a concern for most pre-training setups (Uthus et al., 2023), including those of encoder models. This observation is in line with the study of Nguyen et al. (2024) and indicates that language models are likely more proficient when English is the target language, which consequently leads to enhanced performance by LLMs and encoders, and poses a question on multilingual claims made by LLM releases.

Figure 2 indicates LLMs are outperformed for most language pairs by TransQuest-based and COMET models. Interestingly, for the only language pair where LLMs match COMET performance, En-Mr, the results are statistically insignificant. Among the Indic-target language pairs, En-Mr shows a consistently higher correlation, but statistically insignificant in most cases (Table 2) across both settings. In the UMT setting, this could be an outcome of imbalanced data distribution since En-Mr has a significantly large training set, but we have similarly insignificant outcomes from the ILT setting as well. Our work indicates that LLM-based adapters may not perform as well as encoder-based models. Investigating larger variants may produce better performance but smaller segment-level encoder-based QE models render this direction inefficient. Further, due to the black-box nature of Transformer-based language models, we resort to a tokenization analysis which reveals likely explanations for their QE performance.

Tokenization analysis To explore the reasons behind the better performance of fine-tuned pre-trained encoders over LLMs in reference-less QE tasks, we conducted an analysis of token counts generated by LLMs and pre-trained encoders, such as TransQuest's InfoXLM and CometKiwi's XLM-R-XL. For comparison, high-resource language pairs from the WMT22 test data (Zerva et al., 2022) were included to assess tokenization differences across languages with varying resources.

We selected 100 sentences per language pair from our test set and created a tokenization pipeline for each model. Both source and translation texts were input to observe token counts. Figure 4 shows word and token counts for three language pairs, revealing slight differences between Llama-2-7B and OpenChat-3.5 despite using the same tokenizer. The tokenization outcomes for all language pairs are detailed in Appendix E. The token counts generated by LLMs (Gemma, OpenChat, Llama) for low-resource non-English languages significantly deviate from the original word counts, while pre-trained encoders like InfoXLM and XLMR-XL show smaller discrepancies. Rich morphological languages like Marathi, Tamil, and Telugu, which feature agglutinative[6] phrases, and Hindi, which includes compounding, experience skewed tokenization, affecting semantic matching between source and translation (Appendix E). In contrast, for English, the tokenized count closely matches the word count, regardless of the model used. This highlights the need for improved tokenization strategies for cross-lingual semantic matching with LLMs for low-resource languages to enhance performance on the QE task.

We also identified that the Et-En language pair consistently achieved the highest performance across all experimental settings. As illustrated in the Appendix E, the difference between the token counts generated by language models *vs.* the original word counts is evidently smaller than

[6]A grammatical process in which words are composed of a sequence of morphemes (meaningful word elements), each of which represents not more than a single grammatical category

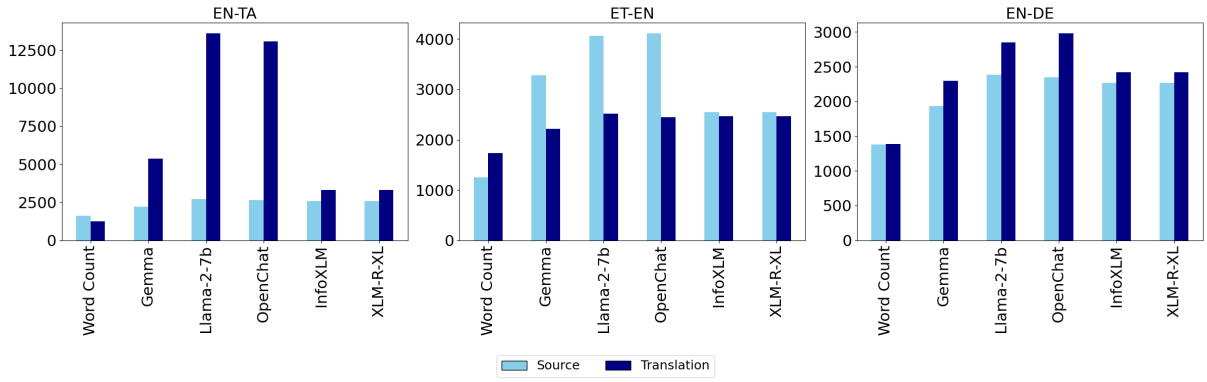

Figure 4: The graphs compare the original word counts with the model-generated token counts for selected inputs, as described in Section 5. This comparison includes both low-resource language pairs (En-Ta, Et-En) and a high-resource language pair (En-De). A detailed image covering all language pairs is provided in the Appendix E.

that observed for other low-resource languages. This holds true even for LLMs as well. This reduced tokenization discrepancy, likely due to both languages (En and Et) using the Latin alphabet, may explain why Et-En performs better in all experimental settings.

Looking at the tokenization for high-resource non-English languages (see Appendix E), it can be seen that the language pairs En-De (German) and Ro (Romanian)-En exhibit limited disparity in the number of tokens from the original word count, and for En-Zh (Chinese) it significantly higher. De/Ro uses Latin-based scripts too. This analysis suggests that English and other Latin-script-based languages, benefit from more efficient tokenization in language models, which leads to improved performance in tasks like QE. In contrast, other languages, exhibit greater disparities in token counts, indicating the need for more advanced tokenization strategies with LLMs to enhance performance. This underscores the importance of developing better tokenization methods to ensure equitable model performance across different language pairs.

Error Analysis We conducted an error analysis using the top-performing model, OpenChat, focusing solely on the En-Ta language pair due to native speaker availability. The purpose of this analysis was to identify the underlying reasons for significant deviations in predicted DA scores from the ground truths, aiming to understand what factors in the input contribute to inaccurate predictions. From the model's predictions, we selected 100 sentences with the highest deviations between predicted and human-annotated DA scores. Figure 3 presents the identified error types and their occurrence percentages. The annotated error types are based on the Multidimensional Quality Metric Error typology (Lommel et al., 2014).

A significant portion of errors, such as *Incorrect term* (26.3%), *Use of Entity* (16.8%), and *Syntactic error* (14.6%), suggests that the model struggles with accurately understanding the contextual appropriateness in the translations to predict the DA score. This can be attributed to the inherent challenges in capturing the nuances and complexities of language, especially for low-resource languages where the training data may be insufficient or lacks diversity. The *Long-text* (13.9%) and *Incomplete sentence* (8%) errors indicate difficulties in maintaining coherence and completeness in translation, which are crucial for accurate QE. *Missing information* (12.4%) which highlights the challenge of ensuring the completeness of the sentence and *Transliteration* errors (2.2%) highlighting the challenges of understanding the conversion of phonetic elements also seem to be important for accurate quality estimation. Finally *Use of abbreviation* errors (4.4%) suggest that the model is unlikely to have seen domain-specific terminology, which requires domain-specific training data for better quality estimation.

6 Conclusion and Future Work

This paper investigates reference-less quality estimation for low-resource language pairs using large language models. We reproduce results with existing SOTA prompts and propose a new AG prompt, which performs best in zero-shot settings. Further experiments with ICL and instruction fine-tuning settings are performed with AG prompt which achieves closer performance with the pre-

trained encoder-based approaches.

Our findings indicate how LLM-based QE can be challenging for morphologically richer languages without much data in the pre-training stage. Based on our findings, we highly recommend the addition of QE datasets to LLM evaluation task suits given the significant cross-lingual challenge posed by this task. We perform a detailed tokenization analysis which highlights that cross-lingual machine understanding for low-resource languages needs to be addressed at the stage of tokenization (Remy et al., 2024), and within pre-training data (Petrov et al., 2024). Additionally, error analysis highlights significant challenges in handling context, syntax, and domain-specific terms, suggesting that further refinement in model training and adaptation is necessary. In the future, we aim to employ regression head-based adapters within the LLM pipeline for QE, eliminating the challenges in the reliability of extracting the scores from the outputs.

7 Limitations

Our results are based on a limited number of LLMs, primarily smaller than 14 billion parameters, due to the constraints imposed by our computational resources. All experiments were conducted using only one GPU (NVIDIA A40 40G), which required significant time for instruction fine-tuning and inference across several language pairs. Additionally, our study was limited to open-source LLMs.

The availability of human-annotated DA scores for low-resource languages is limited to the eight language pairs included in this study and our analysis is constrained to these specific datasets. In the future, we aim to expand our study to include datasets where the source and translated languages are reversed, provided such datasets become available.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Frederic Blain, Chrysoula Zerva, Ricardo Rei, Nuno M. Guerreiro, Diptesh Kanojia, José G. C. de Souza, Beatriz Silva, Tânia Vaz, Yan Jingxuan, Fatemeh Azadi, Constantin Orasan, and André Martins. 2023. Findings of the WMT 2023 shared task on quality estimation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 629–653, Singapore. Association for Computational Linguistics.

Jiuhai Chen, Lichang Chen, Chen Zhu, and Tianyi Zhou. 2023. How many demonstrations do you need for in-context learning? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11149–11159, Singapore. Association for Computational Linguistics.

Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1066–1083, Singapore. Association for Computational Linguistics.

Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. 2021. Larger-scale transformers for multilingual masked language modeling. *CoRR*, abs/2105.00572.

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.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting. *arXiv preprint arXiv:2305.07004*.

Xu Huang, Zhirui Zhang, Xiang Geng, Yichao Du, Jiajun Chen, and Shujian Huang. 2024. Lost in the source language: How large language models evaluate the quality of machine translation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 3546–3562, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Cong-Phap Huynh, Christian Boitet, and Hervé Blanchon. 2008. Sectra_w. 1: an online collaborative system for evaluating, post-editing and presenting mt translation corpora. In *LREC*, volume 8, pages 28–30.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Mirella Lapata. 2006. Automatic evaluation of information ordering: Kendall's tau. *Computational Linguistics*, 32(4):471–484.

Yuang Li, Chang Su, Ming Zhu, Mengyao Piao, Xinglin Lyu, Min Zhang, and Hao Yang. 2023. HW-TSC 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 835–840, Singapore. Association for Computational Linguistics.

Chi-kiu Lo, Meriem Beloucif, Markus Saers, and Dekai Wu. 2014. Xmeant: Better semantic mt evaluation without reference translations. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 765–771.

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

Shushen Manakhimova, Eleftherios Avramidis, Vivien Macketanz, Ekaterina Lapshinova-Koltunski, Sergei Bagdasarov, and Sebastian Möller. 2023. Linguistically motivated evaluation of the 2023 state-of-the-art machine translation: Can ChatGPT outperform NMT? In *Proceedings of the Eighth Conference on Machine Translation*, pages 224–245, Singapore. Association for Computational Linguistics.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. 2018. Mixed precision training. *Preprint*, arXiv:1710.03740.

Vandan Mujadia, Pruthwik Mishra, Arafat Ahsan, and Dipti M. Sharma. 2023. Towards large language model driven reference-less translation evaluation for English and Indian language. In *Proceedings of the*

20th International Conference on Natural Language Processing (ICON), pages 357–369, Goa University, Goa, India. NLP Association of India (NLPAI).

Xuan-Phi Nguyen, Mahani Aljunied, Shafiq Joty, and Lidong Bing. 2024. Democratizing LLMs for low-resource languages by leveraging their English dominant abilities with linguistically-diverse prompts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3501–3516, Bangkok, Thailand. Association for Computational Linguistics.

Aleksandar Petrov, Emanuele La Malfa, Philip Torr, and Adel Bibi. 2024. Language model tokenizers introduce unfairness between languages. *Advances in Neural Information Processing Systems*, 36.

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2020. TransQuest: Translation quality estimation with cross-lingual transformers. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5070–5081, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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

Ricardo Rei, Nuno M. Guerreiro, JosÃ© 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.

François Remy, Pieter Delobelle, Hayastan Avetisyan, Alfiya Khabibullina, Miryam de Lhoneux, and Thomas Demeester. 2024. Trans-tokenization and cross-lingual vocabulary transfers: Language adaptation of llms for low-resource nlp. *arXiv preprint arXiv:2408.04303*.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.

Philip Sedgwick. 2014. Spearman's rank correlation coefficient. *Bmj*, 349.

Archchana Sindhujan, Diptesh Kanojia, Constantin Orasan, and Tharindu Ranasinghe. 2023. SurreyAI 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 849–855, Singapore. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

David Uthus, Santiago Ontanon, Joshua Ainslie, and Mandy Guo. 2023. mLongT5: A multilingual and efficient text-to-text transformer for longer sequences. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9380–9386, Singapore. Association for Computational Linguistics.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2023. Openchat: Advancing open-source language models with mixed-quality data.

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Wang, and Lei Li. 2023. INSTRUCTSCORE: Towards explainable text generation evaluation with automatic feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5967–5994, Singapore. Association for Computational Linguistics.

Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. 2022. Findings of the WMT 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu,

Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *Preprint*, arXiv:2303.18223.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14544–14556, Singapore. Association for Computational Linguistics.

A Appendix: In-context learning prompt

We need to Evaluate the machine translated sentences of **<Source language>(Source)** to **<Target language> (Translation)**, with quality scores ranging from 0 to 100. In following examples shows the quality score given by a human translator.

Example 1 - <An example of a source and translated sentence with the human annotated DA score in the range of **0-30 >**

Example 2 - < An example of a source and translated sentence with the human annotated DA score in the range of **31-50 >**

Example 3 - < An example of a source and translated sentence with the human annotated DA score in the range of **51-70 >**

Example 4 - < An example of a source and translated sentence with the human annotated DA score in the range of **71-90 >**

Example 5 - < An example of a source and translated sentence with the human annotated DA score in the range of **91-100 >**

Scores of 0-30 indicate that the translation is mostly unintelligible, either completely inaccurate or containing only some keywords. Scores of 31-50 suggest partial intelligibility, with some keywords present but numerous grammatical errors. A score between 51-70 means the translation is generally clear, with most keywords included and only minor grammatical errors. Scores of 71-90 indicate the translation is clear and intelligible, with all keywords present and only minor non-grammatical issues. Finally, scores of 91-100 reflect a perfect or near-perfect translation, accurately conveying the source meaning without errors. The evaluation criteria focus on two main aspects: Adequacy (how much information is conveyed) and Fluency (how grammatically correct the translation is).

Predict the quality score for the following translation in the range of 0 to 100, considering the above instructions and given examples. Predict only the score, no need for explanation.

Source: <Source Sentence>

Translation: <Translation Sentence>

Score is

Figure 5: Our proposed AG prompt for in-context learning.

B Appendix: Other prompts

```
Score the following translation from {Source Language} to {Target
Language} on a continuous scale from 0 to 100, where score of zero
means "no meaning preserved" and score of one hundred means "perfect
meaning and grammar".

{Source Language} source: {Source Sentence}

{Target Language} translation: {Translated Sentence}

Score:
```

Figure 6: GEMBA prompt (Kocmi and Federmann, 2023)

The **GEMBA** prompt is part of the GEMBA (GPT Estimation Metric Based Assessment) method, which uses GPT-based language models to evaluate translation quality. The GEMBA prompt evaluates translation quality by scoring each translation segment on a continuous scale from 0 to 100.

```
You are an experienced translation evaluator and you need to evaluate
a translation for <Source Language> language to <Target Language>
language.

{Source Language}: {Source Sentence}

{Target Language}: {Translated Sentence}

The evaluation score out of 100 is
```

Figure 7: TE prompt (Mujadia et al., 2023)

The **TE (Translation Evaluator)** prompt instructs the model to act as an experienced translation evaluator, explicitly presenting the source language, source text, target language, and translated text. The prompt concludes with the model assigning a score out of 100 to the translation, indicating its quality.

C Appendix: Train and test data splits

Lang.	Train	Test
English - Gujarati (En-Gu)	7000	1000
English - Hindi (En-Hi)	7000	1000
English - Marathi (En-Mr)	26 000	699
English - Tamil (En-Ta)	7000	1000
English - Telugu (En-Te)	7000	1000
Estonian - English (Ne-En)	7000	1000
Nepalis - English (Ne-En)	7000	1000
Sinhala - English (Si-En)	7000	1000

Table 3: The dataset splits of translation datasets with human-annotated DA scores utilized in our study. We conducted experiments on 8 low-resource language pairs to evaluate the performance of various models.

D Appendix: Train and test data with number of instances in each DA score ranges

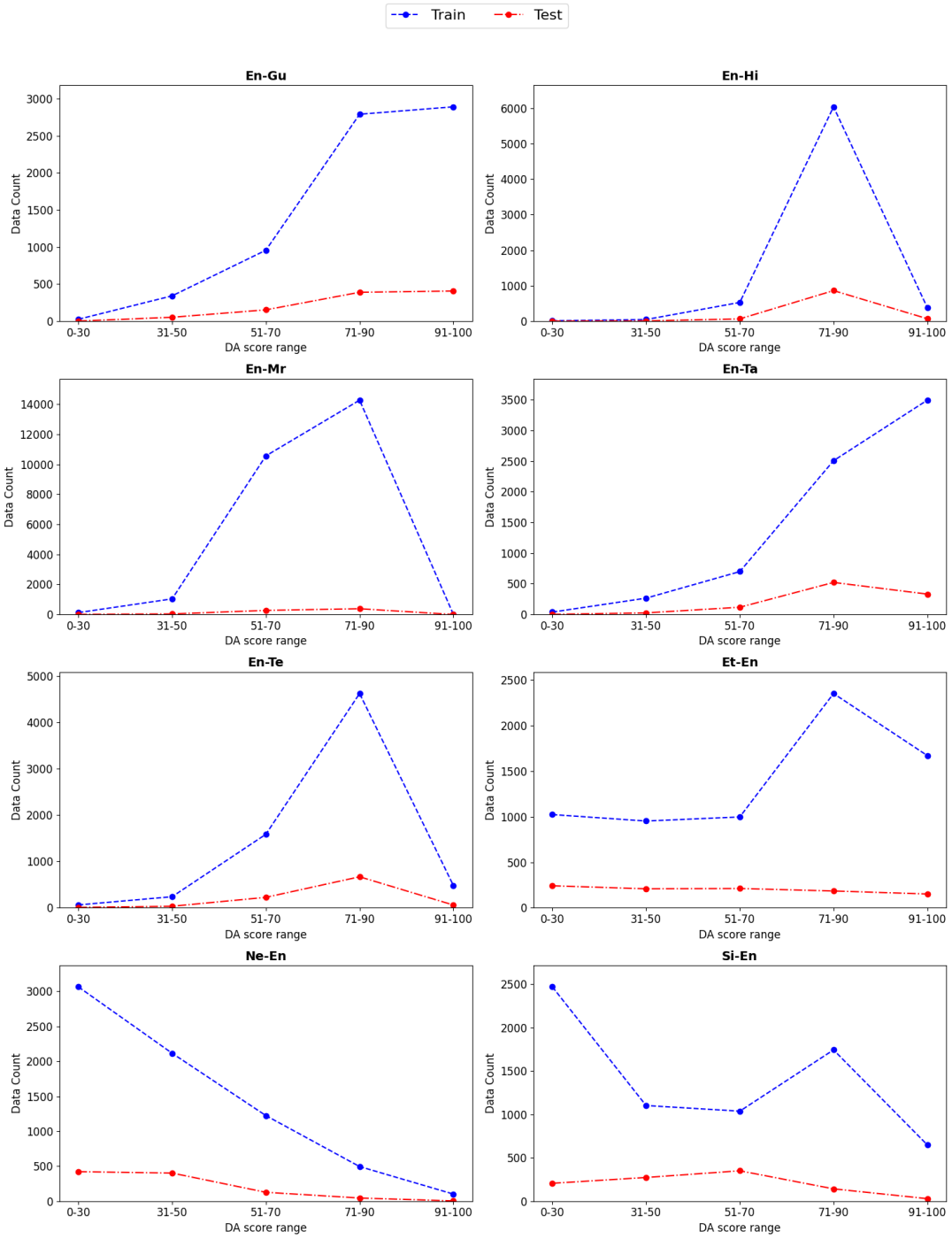

Figure 8: This image shows the number of data belonging to each DA score range of each language pair in the train and test data sets.

E Appendix: Tokenization with different language models

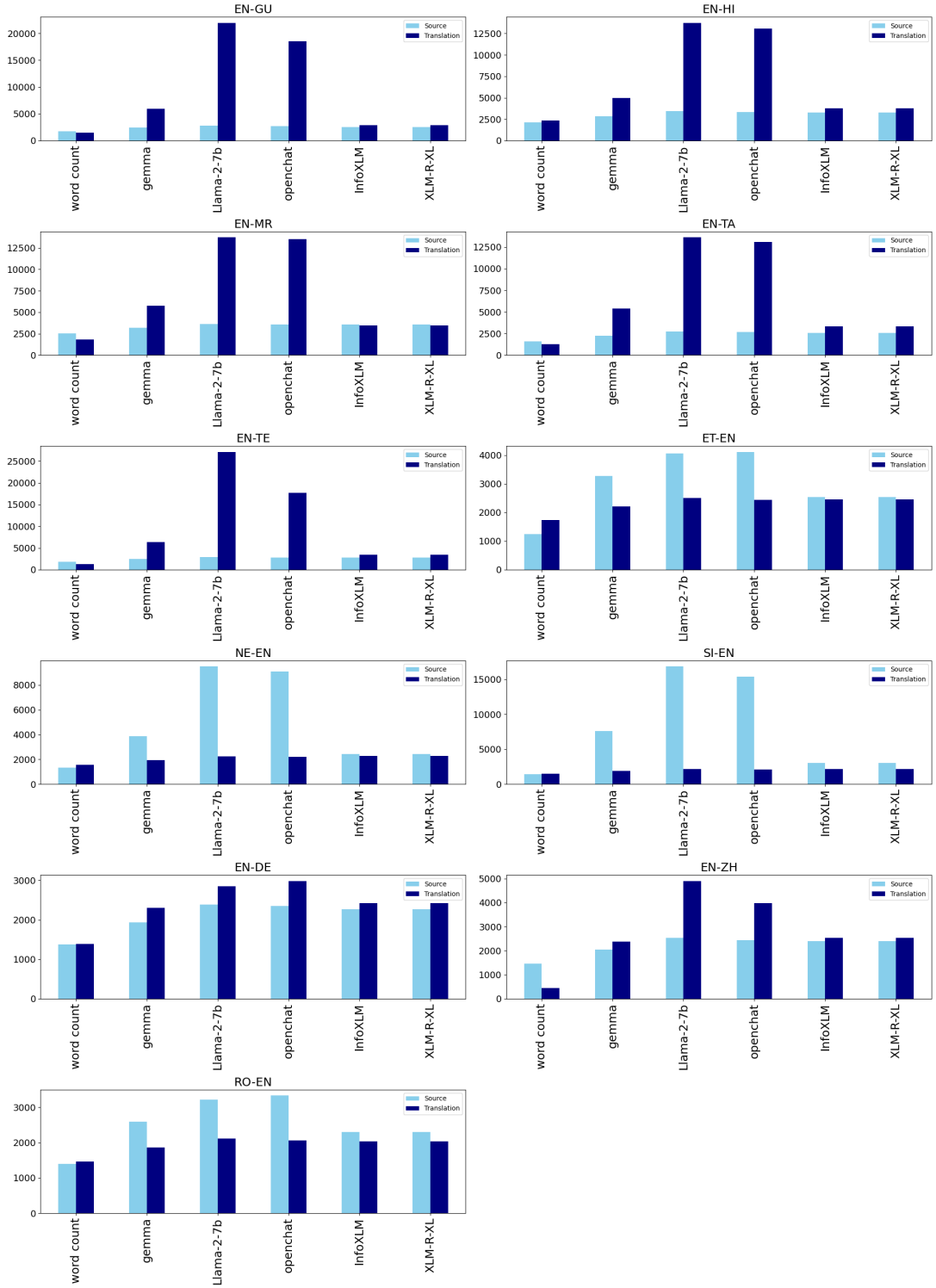

Figure 9: A comparative analysis of the total word count of source and target sentences versus the count of tokens generated by various language models for, both low-resource and high-resource language pairs. The X-axis represents the model name, while the Y-axis indicates the generated token counts.

F Appendix: Zero-shot experiment results with Pearson, Spearman and Kendal's Tau Correlation scores

Language pairs	Gemma-7B				Llama-2-7B				Llama-2-13B				OC-3.5-7B			
	r	ρ	τ	E	r	ρ	τ	E	r	ρ	τ	E	r	ρ	τ	E
En-Gu	0.125	0.113	0.092	0	0.015	0.006	0.005	1	0.048	0.019	0.016	6	0.267	0.249	0.187	1
En-Hi	0.154	0.131	0.106	0	-0.031	-0.002	-0.001	6	0.049	0.009	0.007	6	0.315	0.254	0.188	9
En-Mr	0.177	0.135	0.109	0	0.054	0.053	0.042	0	0.103	0.115	0.088	7	0.323	0.183	0.137	0
En-Ta	0.346	0.222	0.179	0	0.034	0.067	0.054	17	0.108	0.091	0.070	6	0.400	0.358	0.270	4
En-Te	0.074	0.081	0.066	0	0.005	-0.016	-0.013	0	0.093	0.121	0.094	0	0.155	0.145	0.109	0
Et-En	0.286	0.289	0.229	1	0.173	0.168	0.129	3	0.232	0.185	0.139	26	0.550	0.571	0.411	3
Ne-En	0.261	0.261	0.199	1	0.144	0.153	0.119	10	0.234	0.222	0.165	11	0.476	0.448	0.320	14
Si-En	0.272	0.193	0.150	5	0.155	0.144	0.113	7	0.232	0.195	0.146	5	0.439	0.417	0.299	8

Table 4: The complete results of the zero-shot experiments using large language models and the GEMBA prompt template (Kocmi and Federmann, 2023). The results include Pearson (r), Spearman (ρ), and Kendall's Tau (τ) correlation scores. The column 'E' indicates the number of rows excluded because the outputs generated by the large language models did not include a score.

Language pairs	Gemma-7B				Llama-2-7B				Llama-2-13B				OC-3.5-7B			
	r	ρ	τ	E	r	ρ	τ	E	r	ρ	τ	E	r	ρ	τ	E
En-Gu	-0.094	-0.102	-0.085	27	-0.024	-0.008	-0.005	14	-0.045	-0.052	-0.039	102*	0.180	0.117	0.085	50
En-Hi	-0.056	-0.050	-0.041	10	-0.022	-0.072	-0.051	28	0.047	0.056	0.041	40	0.239	0.134	0.095	51
En-Mr	0.209	0.173	0.141	12	0.070	0.070	0.048	20	0.072	0.040	0.030	48	0.192	0.114	0.080	34
En-Ta	-0.017	-0.037	-0.030	17	-0.002	0.012	0.009	47	-0.036	0.016	0.011	143 *	0.178	0.178	0.126	66
En-Te	0.026	0.018	0.015	37	0.026	0.013	0.009	26	-0.007	0.010	0.008	68	0.073	0.072	0.051	59
Et-En	0.098	0.086	0.070	43	0.129	0.100	0.069	2	0.157	0.146	0.107	28	0.464	0.455	0.322	5
Ne-En	0.153	0.155	0.125	90	0.142	0.100	0.070	25	0.062	0.080	0.060	114*	0.358	0.334	0.235	94
Si-En	0.055	0.055	0.045	20	0.134	0.129	0.091	10	0.100	0.109	0.080	45	0.308	0.303	0.211	43

Table 5: The complete results of the zero-shot experiments using large language models and the TE prompt template (Mujadia et al., 2023). The results include Pearson (r), Spearman (ρ), and Kendall's Tau (τ) correlation scores. The column 'E' indicates the number of rows excluded because the outputs generated by the large language models did not include a score. (*) in the column E indicates that more than 10% of the total inferences were dropped, which means the results may be considered not trustworthy.

Language pairs	Gemma-7B				Llama-2-7B				Llama-2-13B				OC-3.5-7B			
	r	ρ	τ	E	r	ρ	τ	E	r	ρ	τ	E	r	ρ	τ	E
En-Gu	-0.034	-0.079	-0.059	2	0.047	-0.007	-0.006	0	-0.033	0.008	0.007	0	0.159	0.164	0.132	2
En-Hi	-0.042	-0.056	-0.041	0	0.021	-0.029	-0.022	0	0.051	0.069	0.055	1	0.303	0.253	0.200	0
En-Mr	0.033	0.027	0.020	3	0.097	0.059	0.046	0	-0.007	0.005	0.004	1	0.340	0.276	0.222	0
En-Ta	0.026	-0.002	0.000	14	0.009	0.055	0.041	0	-0.026	-0.070	-0.057	1	0.367	0.363	0.290	2
En-Te	0.072	0.065	0.048	0	0.064	0.083	0.065	1	0.010	0.045	0.038	0	0.129	0.121	0.095	0
Et-En	0.077	0.098	0.071	4	0.115	0.064	0.049	1	0.304	0.319	0.255	1	0.615	0.619	0.470	1
Ne-En	0.129	0.130	0.096	47	0.178	0.144	0.111	1	0.283	0.303	0.236	1	0.539	0.487	0.370	5
Si-En	0.037	0.042	0.031	14	0.155	0.069	0.056	5	0.267	0.238	0.185	6	0.466	0.441	0.341	8

Table 6: The complete results of the zero-shot experiments using large language models and the AG prompt template. The results include Pearson (r), Spearman (ρ), and Kendall's Tau (τ) correlation scores. The column 'E' indicates the number of rows excluded because the outputs generated by the large language models did not include a score.

G Appendix: In-context learning experiment results with Pearson, Spearman and Kendal's Tau correlation scores

LP	Gemma-7B				Llama-2-7B				Llama-2-13B				OC-3.5-7B			
	r	ρ	τ	E	r	ρ	τ	E	r	ρ	τ	E	r	ρ	τ	E
En-Gu	0.010	-0.005	0.003	26	0.052	0.036	0.028	1	-0.071	-0.036	-0.029	1	0.202	0.223	0.174	0
En-Hi	0.135	0.134	0.097	71	-0.059	-0.114	-0.089	1	0.009	-0.023	-0.019	0	0.237	0.184	0.146	0
En-Mr	0.243	0.202	0.145	111*	0.130	0.120	0.089	2	0.093	0.095	0.069	0	0.249	0.218	0.173	0
En-Ta	0.106	0.122	0.089	81	0.015	-0.019	-0.013	27	0.068	0.083	0.061	1	0.252	0.337	0.270	0
En-Te	0.104	0.092	0.068	53	0.038	0.027	0.021	25	-0.001	0.015	0.012	0	0.083	0.152	0.124	0
Et-En	0.233	0.226	0.162	14	0.268	0.268	0.198	9	0.009	-0.058	-0.048	4	0.590	0.613	0.459	1
Ne-En	0.275	0.273	0.195	78	0.161	0.149	0.110	5	0.322	0.340	0.266	1	0.486	0.457	0.346	1
Si-En	0.312	0.306	0.219	56	0.158	0.146	0.109	19	0.150	0.018	0.013	5	0.484	0.470	0.348	5

Table 7: The complete results of the ICL experiment with 3 examples using our proposed AG prompt template (3-ICL-AG). The results include Pearson (r), Spearman (ρ), and Kendall's Tau (τ) correlation scores. 'LP'-> Language Pair, 'E'-> the number of rows excluded because the outputs generated by the large language models did not include a score. (*) in the column E indicates that more than 10% of the total inferences were dropped, which means the results may be considered not trustworthy.

LP	Gemma-7B				Llama-2-7B				Llama-2-13B				OC-3.5-7B			
	r	ρ	τ	E	r	ρ	τ	E	r	ρ	τ	E	r	ρ	τ	E
En-Gu	0.008	0.023	0.016	68	0.002	-0.008	-0.006	1	0.087	0.095	0.070	0	0.157	0.151	0.120	0
En-Hi	0.134	0.075	0.054	32	0.002	-0.022	-0.016	0	0.031	0.035	0.027	0	0.243	0.212	0.163	0
En-Mr	0.218	0.164	0.119	25	0.035	0.032	0.023	0	0.028	-0.031	-0.026	0	0.256	0.226	0.181	0
En-Ta	0.099	0.114	0.081	92	-0.010	0.017	0.013	1	0.095	0.193	0.146	0	0.324	0.332	0.263	0
En-Te	0.006	0.021	0.015	91	0.067	0.051	0.038	0	0.023	0.073	0.057	0	0.075	0.126	0.101	0
Et-En	0.318	0.327	0.231	86	0.263	0.269	0.194	2	0.461	0.438	0.322	1	0.604	0.636	0.482	1
Ne-En	0.311	0.305	0.218	98	0.203	0.189	0.138	3	0.336	0.319	0.243	1	0.502	0.471	0.352	1
Si-En	0.322	0.320	0.230	37	0.123	0.243	0.186	7	0.380	0.326	0.252	5	0.481	0.479	0.358	5

Table 8: The complete results of the ICL experiment with 5 examples using our proposed AG prompt template (5-ICL-AG). The results include Pearson (r), Spearman (ρ), and Kendall's Tau (τ) correlation scores. 'LP'-> Language Pair, 'E'-> the number of rows excluded because the outputs generated by the large language models did not include a score.

LP	Gemma-7B				Llama-2-7B				Llama-2-13B				OC-3.5-7B-1210			
	r	ρ	τ	E	r	ρ	τ	E	r	ρ	τ	E	r	ρ	τ	E
En-Gu	0.060	0.071	0.052	62	0.022	-0.053	-0.043	3	-0.093	-0.108	-0.082	0	0.222	0.260	0.203	0
En-Hi	0.116	0.075	0.053	64	-0.088	-0.176	-0.139	2	0.045	0.014	0.011	0	0.173	0.163	0.128	0
En-Mr	0.256	0.167	0.126	50	0.075	0.050	0.040	5	0.068	0.047	0.036	0	0.277	0.251	0.201	0
En-Ta	0.094	0.122	0.086	80	-0.083	-0.096	-0.074	0	-0.059	-0.004	-0.004	0	0.285	0.309	0.233	0
En-Te	-0.039	-0.033	-0.025	51	0.044	0.021	0.016	0	-0.009	-0.028	-0.023	0	0.095	0.196	0.149	1
Et-En	0.305	0.306	0.218	39	0.052	0.033	0.025	1	0.198	0.169	0.125	1	0.595	0.616	0.469	1
Ne-En	0.363	0.365	0.263	86	-0.009	-0.040	-0.032	1	0.215	0.259	0.210	1	0.511	0.491	0.374	1
Si-En	0.284	0.283	0.203	33	-0.019	-0.017	-0.011	5	0.287	0.223	0.164	5	0.462	0.477	0.351	5

Table 9: The complete results of the ICL experiment with 7 examples using our proposed AG prompt template (7-ICL-AG). The results include Pearson (r), Spearman (ρ), and Kendall's Tau (τ) correlation scores. 'LP'-> Language Pair, 'E'-> the number of rows excluded because the outputs generated by the large language models did not include a score.

H Appendix: Complete results of unified multilingual training based fine-tuned experiments

LP	Gemma-7B			Llama-2-7B			Llama-2-13B			OC-3.5-7B			TransQuest			CometKiwi		
	r	ρ	τ	r	ρ	τ	r	ρ	τ	r	ρ	τ	r	ρ	τ	r	ρ	τ
En-Gu	0.628	0.566	0.424	0.551	0.461	0.339	0.558	0.465	0.345	0.616	0.554	0.418	0.680	0.630	0.460	0.678	0.637	0.467
En-Hi	0.570	0.449	0.333	0.490	0.332	0.242	0.486	0.322	0.235	0.585	0.458	0.341	0.610	0.478	0.336	0.648	0.615	0.446
En-Mr	0.631	0.551	0.401	0.573	0.516	0.376	0.589	0.505	0.369	0.631	0.545	0.397	0.658	0.606	0.434	0.618	0.546	0.390
En-Ta	0.584	0.502	0.382	0.488	0.464	0.341	0.533	0.471	0.351	0.548	0.509	0.385	0.650	0.603	0.435	0.711	0.635	0.455
En-Te	0.179	0.242	0.175	0.228	0.258	0.188	0.227	0.258	0.190	0.211	0.267	0.195	0.330	0.358	0.247	0.310	0.338	0.235
Et-En	0.688	0.728	0.534	0.594	0.636	0.455	0.622	0.655	0.469	0.643	0.678	0.493	0.755	0.760	0.560	0.853	0.860	0.661
Ne-En	0.688	0.650	0.476	0.598	0.519	0.370	0.628	0.565	0.404	0.657	0.607	0.438	0.767	0.718	0.530	0.783	0.789	0.599
Si-En	0.469	0.455	0.320	0.408	0.395	0.275	0.410	0.403	0.281	0.489	0.481	0.339	0.627	0.579	0.413	0.730	0.703	0.515

Table 10: The complete results of the UMT instruction fine-tuning experiment with large language models and pre-trained encoder-based approaches (TransQuest-InfoXLM, CometKiwi-XLM-R-XL) for low-resourced language pairs (LP). The results include Pearson (r), Spearman (ρ), and Kendall's Tau (τ) correlation scores.

I Appendix: Complete results of independent language-pair training based fine-tuned experiments

LP	Gemma-7B			Llama-2-7B			Llama-2-13B			OC-3.5-7B			TransQuest		
	r	ρ	τ	r	ρ	τ	r	ρ	τ	r	ρ	τ	r	ρ	τ
En-Gu	0.531	0.440	0.326	0.189	0.214	0.153	0.463	0.421	0.311	0.583	0.520	0.388	0.690	0.653	0.477
En-Hi	0.482	0.375	0.276	0.317	0.282	0.204	0.406	0.336	0.247	0.575	0.474	0.354	0.134	0.119	0.080
En-Mr	0.617	0.557	0.407	0.548	0.509	0.371	0.555	0.501	0.364	0.630	0.554	0.406	0.508	0.629	0.447
En-Ta	0.544	0.475	0.355	0.398	0.375	0.274	0.459	0.441	0.326	0.551	0.509	0.379	0.268	0.303	0.205
En-Te	0.135	0.217	0.155	0.202	0.263	0.193	0.202	0.261	0.191	0.211	0.271	0.199	0.079	0.087	0.059
Et-En	0.622	0.648	0.467	0.569	0.589	0.417	0.559	0.598	0.421	0.609	0.652	0.470	0.797	0.806	0.603
Ne-En	0.660	0.612	0.441	0.545	0.497	0.352	0.582	0.543	0.388	0.646	0.614	0.444	0.777	0.746	0.554
Si-En	0.402	0.387	0.269	0.351	0.332	0.230	0.366	0.346	0.240	0.456	0.441	0.310	0.619	0.581	0.414

Table 11: The complete results of the ILT instruction fine-tuning experiment with large language models and pre-trained encoder-based approach (TransQuest-InfoXLM) for low-resourced language pairs (LP). The results include Pearson (r), Spearman (ρ), and Kendall's Tau (τ) correlation scores.

J Appendix: Examples from error analysis of English-Tamil translation QE task

Source and Translated Sentences	Error Description
Source: Aitzaz Hasan's father is Mujahid Ali, who was in the United Arab Emirates at the time of the attack. **Translation:** தாக்குதலின் போது ஐக்கிய அரபு எமிரேட்ஸில் இருந்த ஐத்ஸாஸ் ஹசனின் தந்தை முஜாஹித் அலி.	Use of Entity – This sentence contains many named entities. Syntactic error - The structure of the sentence is not correctly presented.
Source: But the internship.. **Translation:** ஆனால் உள்ளக...	Incomplete sentence – The source sentence is an incomplete sentence which led to the incomplete translation.
Source: What does she burn? **Translation:** அவளுக்கு என்ன எரிச்சல்?	Incorrect term – The word-for-word translation of "burn" is accurate, but the sentence fails to convey the intended meaning. The translation misinterprets the context, resulting in an incorrect overall interpretation of the original sentence.
Source: Yes, I will be your Mom. **Translation:** ஆமாம், நான் உங்கள் தாய் இருப்பேன்.	Syntactic error – The current translation incorrectly maintains the structure of the source sentence, making it unnatural in Tamil syntax.
Source: PAT GOLD grins. **Translation:** PAT தங்க முணுமுணுப்புகள்.	Use of abbreviation – The abbreviation "PAT" should be transliterated or translated in a way that retains its meaning in the context.
Source: Mahfouz's mother, Fatimah, was the daughter of Mustafa Qasheesha, an Al-Azhar sheikh, and although illiterate herself, took the boy Mahfouz on numerous excursions to cultural locations such as the Egyptian Museum and the Pyramids. **Translation:** மகபூஸின் தாயார், பாத்திமா, அல்-அசார் ஷேக் முஸ்தபா காஷீஷாவின் மகள் ஆவார், மேலும் படிப்பறிவு இல்லாதபோதிலும், மகபூஸை எகிப்திய அருங்காட்சியகம் மற்றும் பிரமிடுகள் போன்ற கலாச்சார இடங்களுக்கு ஏராளமான சுற்றுலாக்களில் அழைத்துச் சென்றார்.	Use of Entity and Long-text – This sentence contains many name entities, and the text is longer.
Source: After World War II train services resumed and a steady pattern of service developed at Saltwood, seeing it outlive many of its contemporaries. **Translation:** இரண்டாம் உலகப் போருக்குப் பிறகு ரயில் சேவைகள் மீண்டும் தொடங்கப்பட்டு, சால்ட்வுட்டில் நிலையான சேவை முறை உருவாக்கப்பட்டது	Missing information and Incomplete sentence – The translation omits significant details from the original sentence.

Figure 10: The examples are taken from our study (See in section 5) analyzing the causes of errors leading to high deviations between human-annotated and predicted DA scores from the best-performing LLM OpenChat for English-Tamil language pair. The words highlighted in red indicate the specific terms causing these errors.

K Appendix: Comparative analysis of results from LLMs in different experimental settings

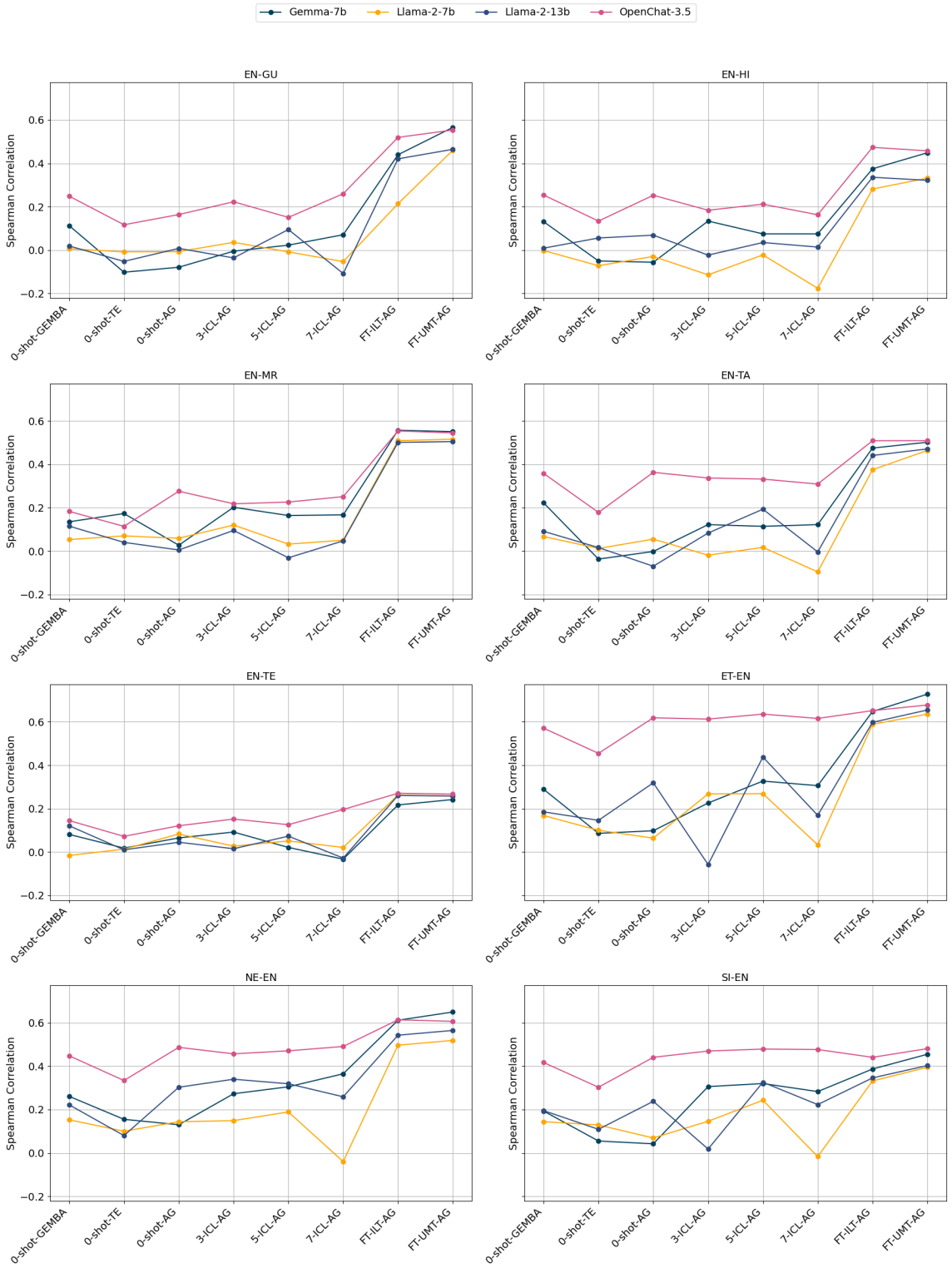

Figure 11: The above graphs show how the Spearman scores varied for each experimental setting with different LLMs. 0-shot-{ GEMBA, TE, AG }-> Zero-shot setting with GEMBA, TE and AG prompts;{N}-ICL-AG -> In-Context-Learning with N number of examples (N = 3, 5, 7) using AG prompt; FT- {ILT, UMT}-AG -> Fine-Tuning with the ILT and UMT setting with the AG prompt.

L Appendix: Models, size and disk space utilization

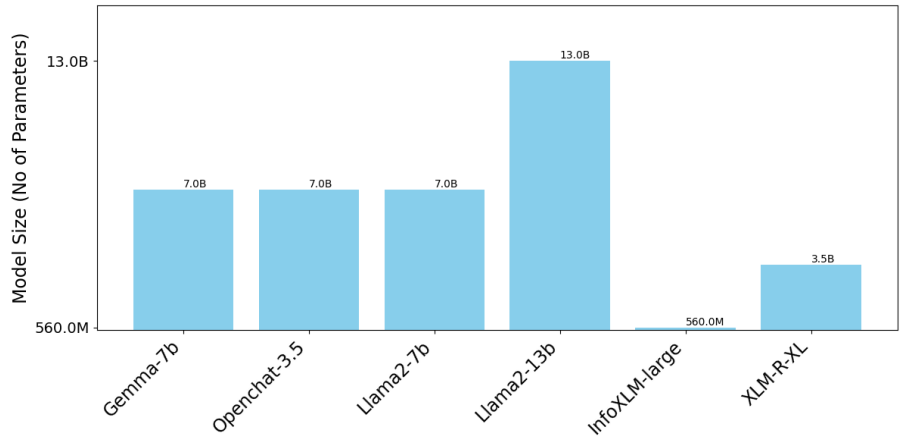

Figure 12: This bar graph shows the size (number of parameters) of the large language models we have utilized for our experiments

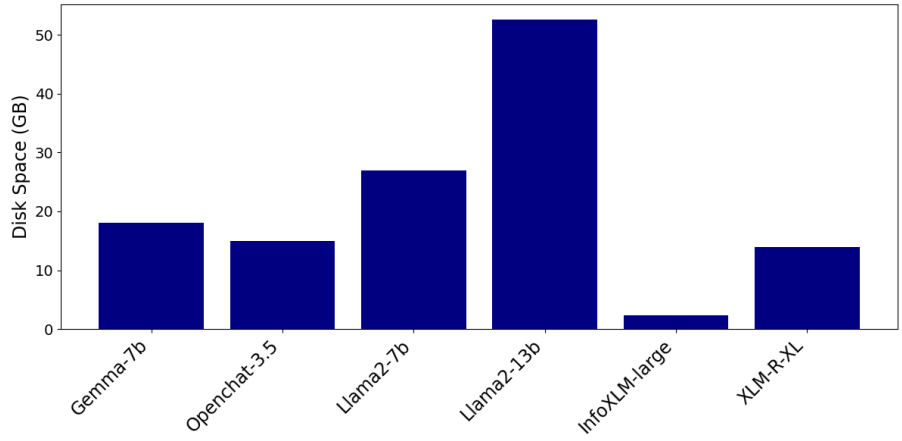

Figure 13: This bar graph shows the disk space utilization of the large language models we have utilized for our experiments

M Appendix: Our publicly available Hugging Face models

Model	Model Link
Gemma-7B	ArchSid/AG-Gemma-7B
Llama-2-7b	ArchSid/AG-Llama-2-7b
Llama-2-13b	ArchSid/AG-Llama-2-13b
Openchat	ArchSid/AG-openchat

Table 12: This table shows the links to our Hugging Face models trained using the Unified Multilingual Training setting.

Model	Language-Pair	Model Link
Gemma-7B	En-Gu	ArchSid/En-Gu_Mono-AG-Gemma-7b
	En-Hi	ArchSid/En-Hi_Mono-AG-Gemma-7b
	En-Mr	ArchSid/En-Mr_Mono-AG-Gemma-7b
	En-Ta	ArchSid/En-Ta_Mono-AG-Gemma-7b
	En-Te	ArchSid/En-Te_Mono-AG-Gemma-7b
	Et-En	ArchSid/Et-En_Mono-AG-Gemma-7b
	Ne-En	ArchSid/Ne-En_Mono-AG-Gemma-7b
	Si-En	ArchSid/Si-En_Mono-AG-Gemma-7b
Llama-2-7b	En-Gu	ArchSid/En-Gu_Mono-AG-Llama-2-7b
	En-Hi	ArchSid/En-Hi_Mono-AG-Llama-2-7b
	En-Mr	ArchSid/En-Mr_Mono-AG-Llama-2-7b
	En-Ta	ArchSid/En-Ta_Mono-AG-Llama-2-7b
	En-Te	ArchSid/En-Te_Mono-AG-Llama-2-7b
	Et-En	ArchSid/Et-En_Mono-AG-Llama-2-7b
	Ne-En	ArchSid/Ne-En_Mono-AG-Llama-2-7b
	Si-En	ArchSid/Si-En_Mono-AG-Llama-2-7b
Llama-2-13b	En-Gu	ArchSid/En-Gu_Mono-AG-Llama-2-13b
	En-Hi	ArchSid/En-Hi_Mono-AG-Llama-2-13b
	En-Mr	ArchSid/En-Mr_Mono-AG-Llama-2-13b
	En-Ta	ArchSid/En-Ta_Mono-AG-Llama-2-13b
	En-Te	ArchSid/En-Te_Mono-AG-Llama-2-13b
	Et-En	ArchSid/Et-En_Mono-AG-Llama-2-13b
	Ne-En	ArchSid/Ne-En_Mono-AG-Llama-2-13b
	Si-En	ArchSid/Si-En_Mono-AG-Llama-2-13b
OpenChat	En-Gu	ArchSid/En-Gu_Mono-AG-openchat
	En-Hi	ArchSid/En-Hi_Mono-AG-openchat
	En-Mr	ArchSid/En-Mr_Mono-AG-openchat
	En-Ta	ArchSid/En-Ta_Mono-AG-openchat
	En-Te	ArchSid/En-Te_Mono-AG-openchat
	Et-En	ArchSid/Et-En_Mono-AG-openchat
	Ne-En	ArchSid/Ne-En_Mono-AG-openchat
	Si-En	ArchSid/Si-En_Mono-AG-openchat

Table 13: This table shows the links to our Hugging Face models trained using the Independent Language-Pair training setting.