ALOPE: Adaptive Layer Optimization for Translation Quality Estimation using Large Language Models

Archchana Sindhujan, **Shenbin Qian**, **Chan Chi Chun Matthew**, **Constantin Orăsan** & **Diptesh Kanojia**

Institute for People-Centred AI and Centre for Translation Studies,
School of Computer Science and Electronic Engineering,
University of Surrey, United Kingdom

{a.sindhujan, s.qian, c.orasan, d.kanojia}@surrey.ac.uk,
matthewchancc@gmail.com

Abstract

Large Language Models (LLMs) have shown remarkable performance across a wide range of natural language processing tasks. Quality Estimation (QE) for Machine Translation (MT), which assesses the quality of a source-target pair without relying on reference translations, remains a challenging cross-lingual task for LLMs. The challenges stem from the inherent limitations of existing LLM-based QE systems, which are pre-trained for causal language modelling rather than regression-specific tasks, further elevated by the presence of low-resource languages given pre-training data distribution. This paper introduces ALOPE, an adaptive layer-optimization framework designed to enhance LLM-based QE by restructuring Transformer representations through layer-wise adaptation for improved regression-based prediction. Our framework integrates low-rank adapters (LoRA) with *regression task heads*, leveraging selected pre-trained Transformer layers for improved cross-lingual alignment. In addition to the layer-specific adaptation, ALOPE introduces two strategies—*dynamic weighting*, which adaptively combines representations from multiple layers, and *multi-head regression*, which aggregates regression losses from multiple heads for QE. Our framework shows improvements over various existing LLM-based QE approaches. Empirical evidence suggests that intermediate Transformer layers in LLMs provide contextual representations that are more aligned with the cross-lingual nature of the QE task. We make resultant models and framework code publicly available[1] for further research, also allowing existing LLM-based MT frameworks to be scaled with QE capabilities.

1 Introduction

Quality Estimation (QE) allows the output of Machine Translation (MT) to be evaluated at scale without the need for reference translations (Zerva et al., 2022). MT QE can be performed at different levels, but this paper focuses on segment-level evaluation for low-resource languages, predicting the Direct Assessment (DA) score ($0 \leq x \leq 100$) for each segment (Graham et al., 2013). The ground truth score is obtained by averaging scores assigned by multiple human annotators ($>= 3$) following annotation guidelines that emphasize adequacy, fluency, and *cross-lingual meaning transfer* within source and machine-translated output.

[1] https://github.com/surrey-nlp/ALOPE

Although QE annotation frameworks such as Multi-dimensional Quality Metrics (MQM) (Lommel et al., 2013) provide fine-grained insights into translation errors, they impose a considerable cognitive burden on annotators, often making them more labour-intensive. In contrast, DA scores offer a more efficient alternative for complementing such frameworks in the overall evaluation of segment-level MT quality. Furthermore, DA scores can be leveraged to guide LLMs in reasoning over translation errors and in identifying translations that require human inspection or post-editing.

While Large Language Models (LLMs) exhibit strong performance in translation evaluation when reference translations are available (Kocmi & Federmann, 2023; Qian et al., 2024), assessing cross-lingual meaning transfer in a reference-less setting remains a significant challenge (Mujadia et al., 2023; Sindhujan et al., 2025) for LLMs. Due to the limited representation of languages in pre-training data (Zhu et al., 2024), complexity of translation errors (Sindhujan et al., 2024), and cross-lingual alignment within LLMs (Nguyen et al., 2024), LLM-based QE does not perform as well as state-of-the-art (SoTA) encoder-based models like CometKiwi (Rei et al., 2023).

The prediction of DA scores for QE constitutes a regression task, requiring fine-grained numerical predictions. However, instructing LLMs to predict a DA score through prompt engineering or instruction tuning for the generation task lacks regression-specific training objectives. This limitation often results in outputs with reduced granularity, thereby constraining their effectiveness in tasks like QE that demand precise scalar predictions (Zerva et al., 2024). Furthermore, current LLM-based methods for cross-lingual tasks such as QE predominantly rely on representations from only the final Transformer layer, overlooking potentially valuable intermediate layer embeddings. Recent findings by Kargaran et al. (2025) indicate that optimal cross-lingual alignment is not necessarily achieved at the final Transformer layer for low-resource languages.

To address these challenges, we introduce the Adaptive Layer Optimization for Translation Quality Estimation (ALOPE) framework, a novel approach that integrates dedicated regression heads and low-rank adapters (LoRA) within Transformer layers, facilitating efficient instruction-based fine-tuning with LLMs. Our framework systematically investigates Transformer layers with regression heads, enabling the identification of optimal configurations that maximize cross-lingual transfer learning for segment-level translation quality estimation. Beyond investigating layer-wise embeddings, ALOPE introduces two additional strategies—dynamic weighting and multi-head regression, which leverage multiple Transformer layers for QE. Moreover, the framework is developed to be model-agnostic, enabling seamless integration of any pre-trained LLMs with minimal configuration, thus promoting ease of deployment and scalability.

We evaluate ALOPE by comparing its performance against zero-shot inference and standard instruction fine-tuned (SIFT) LLMs across four open-source models ($\leq$8B parameters) on eight low-resource language pairs. The results highlight that ALOPE consistently outperforms SIFT, demonstrating its effectiveness by identifying optimal Transformer layers with better cross-lingual transfer learning, which benefits the quality estimation. The main contributions of this paper are,

- We propose ALOPE, an adaptable framework that integrates regression heads with LoRA, enabling efficient QE with any LLMs by facilitating optimal cross-lingual representations.
- ALOPE consistently outperforms the best results achieved by standard instruction fine-tuned LLMs across all eight low-resource language pairs evaluated.
- ALOPE introduces two additional approaches, dynamic-weighting and multi-layer regression, which utilize combined Transformer layers strategies and outperform the LLMs baseline under most conditions evaluated.

2 Background

In recent years, significant advancements have been made in the field of quality estimation. The development of QE models has evolved from early feature engineering-based

approaches (Specia et al., 2013; 2015) to more sophisticated methodologies, progressing through basic neural networks (Kim et al., 2017) and deep neural network architectures (Ive et al., 2018; Kepler et al., 2019). Currently, Transformer-based architectures represent the state-of-the-art in QE (Blain et al., 2023; Rei et al., 2022; Ranasinghe et al., 2020; Perrella et al., 2022; Moura et al., 2020; Baek et al., 2020).

Recent advancements have positioned LLMs as a powerful tool across various NLP tasks (Zhao et al., 2023). In the context of translation evaluation, Kocmi & Federmann (2023) introduced the GEMBA metric, using zero-shot prompting across nine GPT variants and four prompt types for three language pairs. While GEMBA achieved best results with reference translations, its performance in referenceless QE remained below SOTA, highlighting the continued challenges LLMs face in reference-free evaluation settings. Mujadia et al. (2023) investigated QE for low-resource languages by pre-tuning adapters on a large English-Indic parallel corpus for machine translation and subsequently fine-tuning the model with supervised QE data. Their findings suggest that pre-tuning with machine translation data does not enhance QE performance for low-resource languages and that LLMs struggle to transfer MT knowledge effectively for QE tasks in these settings.

It has been noted that LLM-based approaches continue to underperform compared to traditional predictor-estimator architectures in QE, particularly for sentence-level scoring (Zerva et al., 2024). Predictor-estimator models, which explicitly treat QE as a regression task, provide finer-grained quality assessments by effectively distinguishing between varying translation qualities (Kim et al., 2017). In contrast, LLMs, which rely on prompt engineering or instruction tuning, tend to produce outputs with limited granularity, often defaulting to narrower scoring ranges. This lack of differentiation restricts their ability to model nuanced quality variations across translations, contributing to their comparatively weaker performance in QE (Kocmi & Federmann, 2023; Vu et al., 2024).

Cross-lingual transfer learning is intended to improve low-resource language performance by transferring knowledge from high-resource languages through shared linguistic features (Choenni et al., 2023; Pham et al., 2024). However, despite its promise, existing studies show that the approach has so far been more effective in high-resource settings (Zhu et al., 2024), while its benefits for low-resource languages remain limited due to data scarcity and linguistic diversity (Zhu et al., 2024). The work of Sindhujan et al. (2025) reinforces the cross-lingual challenge in QE with LLMs, highlighting the poor tokenization issue for low-resource languages. Nonetheless, optimizing cross-lingual transfer remains essential for improving LLM-based QE.

3 Methodology

3.1 Dataset

Our study mainly focuses on the QE dataset, which is derived from the WMT QE shared task (Zerva et al., 2022; Blain et al., 2023). It contains source and translation pairs with human-annotated DA scores[2], which measure the quality of the machine translations. The QE dataset spans eight low-resource language pairs: English to {Gujarati, Hindi, Marathi, Tamil, Telugu}, and {Nepali, Estonian, Sinhala} to English. Hindi and Estonian, while considered mid-resource languages for machine translation (Nguyen et al., 2024), have limited resources available for quality estimation. We utilize the dataset as train and test splits, as detailed in Appendix A.

3.2 Experimental settings

We utilized the GEMBA prompt, which is a simple and straightforward prompt proposed by Kocmi & Federmann (2023) for estimating the quality of the machine-translated text as shown in Figure 1. All our experiments were conducted with open-source generative LLMs which have 8B parameters or less: LLaMA2-7B, LLaMA3.1-8B, LLaMA3.2-3B and Aya-

[2]Mean DA score annotations for Indic-language pair test sets can be shared upon request.

```
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 1: GEMBA prompt for quality estimation (Kocmi & Federmann, 2023)

expanse-8B [3]. Our experiments were conducted under three settings: zero-shot, standard instruction fine-tuning (SIFT) and ALOPE framework. In zero-shot experiments, we obtain the results directly from the pre-trained LLMs using vLLM framework (Kwon et al., 2023) and we have utilized the LLaMA-Factory framework (Zheng et al., 2024) to perform SIFT.

All instruction-fine-tuning experiments (SIFT and ALOPE) with QE data were conducted using integrated multilingual language-pair training, which involved combining training data from the eight low-resource language pairs mentioned in section 3.1. Inference was performed on language-specific test sets to evaluate model performance.

3.3 ALOPE framework

LLMs are optimized for next-token prediction, excelling at generative tasks but struggling with regression-based goals like quality estimation (Sindhujan et al., 2025). This limits their ability to capture fine-grained relationships between language features and numerical scores. Although LLMs refine context through multiple Transformer layers, standard practice usually predicts with the final layer to obtain the outputs.

Addressing these challenges, ALOPE enables LLMs to perform regression by integrating a regression head within the low-rank adapted Transformer architecture. We analyze the impact of adaptive regression heads to determine which Transformer layers contribute most to enhancing the performance of LLMs for cross-lingual QE tasks with low-resource language pairs.

We employed LoRA (Low-Rank Adaptation) (Hu et al., 2022) for efficient fine-tuning, targeting the projection layers of the Transformer architecture. Rather than updating the full parameter set during fine-tuning, we introduce trainable low-rank matrices into the attention mechanism, allowing the model to instruction fine-tune for the QE task while keeping the original weights frozen as shown in Figure 2. The adaptation to a weight matrix W is represented as $W' = W + BA$, where $A \in \mathbb{R}^{r \times d}$ and $B \in \mathbb{R}^{d \times r}$ are the trainable low-rank matrices and $r \ll d$, where r (the rank) is much smaller than the dimension d which ensures a significant reduction in the number of tunable parameters. In our experiments, the rank r is set to 32. To further optimize memory usage and computational efficiency, we combine LoRA with 4-bit quantization (Dettmers et al., 2023). To enable regression-based prediction with LLMs, the ALOPE framework assigns a dedicated regression head integrated into the Transformer architecture. This component is designed to extract hidden representations from any selected LoRA-adapted Transformer layer, thereby facilitating quality estimation using layer-specific contextual embeddings. The hidden state extracted from the final token of the sequence from the chosen LoRA-adapted Transformer layer is passed through a regression head as shown in Figure 2. Let h_k be the hidden state of the final token:

$$\mathbf{h}_k = \mathbf{H}_k[-1] \in \mathbb{R}^d$$

where k is the chosen Transformer layer, d is the dimensionality of the hidden state, $\mathbf{H}_k$ represents the sequence of hidden states at layer k and $[-1]$ represents the index of the

[3] LLaMa-2-7B, LLAMA3.1-8B, LLAMA3.2-3B, Aya-expanse-8B

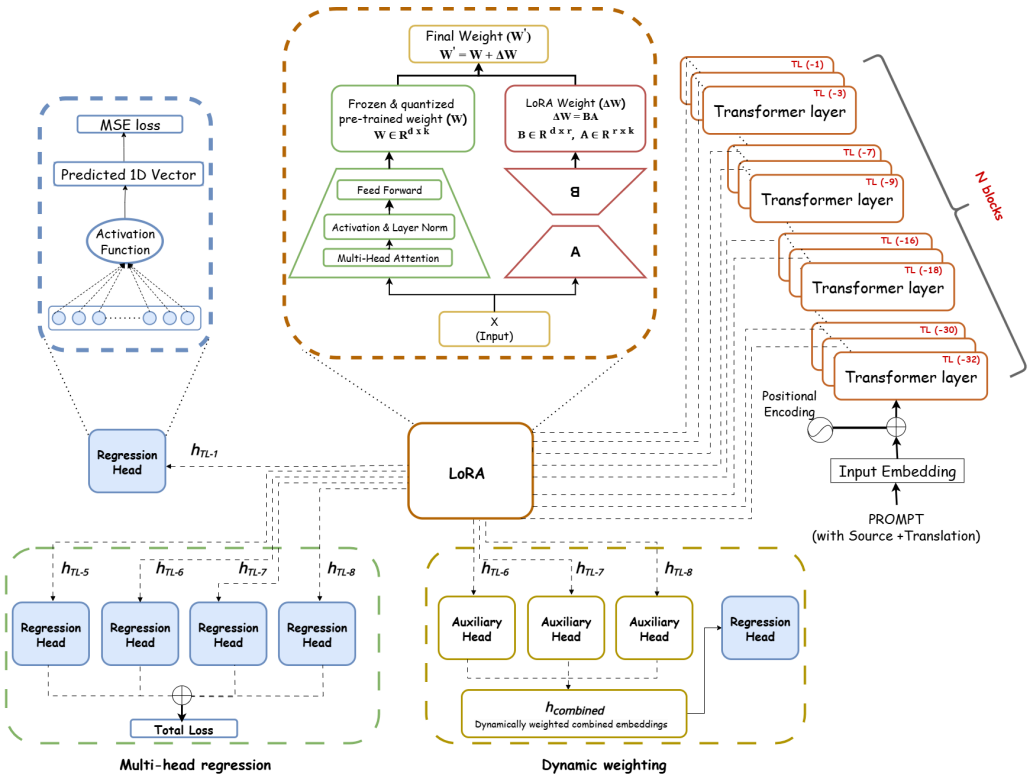

Figure 2: Presents the ALOPE framework. Regression heads can obtain the embeddings from any LoRA-adapted Transformer layers, enabling flexible adaptation to perform QE.

final token in the sequence at Transformer layer k. The regression head consists of a single linear layer that maps the high-dimensional hidden state h_k to the final scalar output. The linear activation function is applied implicitly, as the linear layer performs an affine transformation:

$$\hat{y} = h_k W_n, \quad \hat{y} \in \mathbb{R}$$

where $W_n \in \mathbb{R}^{d \times 1}$ is the weight matrix of the linear layer. The output $\hat{y}$ is a single scalar value per sequence, representing the predicted quality score. To evaluate the model, the mean squared error (MSE) loss is computed by measuring the average squared difference between the predicted values in the 1D vector and the actual target values.

The regression-specific adaptation with regression heads and LoRA-based fine-tuning can be considered independently. However, the focus of our proposed method is efficiency in computationally resource-constrained scenarios. We specifically chose regression heads incorporated with LoRA for its parameter efficiency, reducing the computational overhead compared to full fine-tuning, and allowing modularity given a pre-hosted LLM. ALOPE is able to perform QE with LoRA-less regression heads too. A publicly available library was adopted for our implementation [4].

In the ALOPE framework, the regression heads are assigned numerical indices, where negative values indicate Transformer layers indexed in reverse from the final layer. For example, as shown in the Figure 2, placing a regression head at layer -1 (TL-1) corresponds to extracting the hidden state from the final Transformer layer, while layer-4 (TL-4) refers to the layer located four positions before the last. This indexing approach provides a consistent reference for selecting layers across different models with different numbers of Transformer layers. To identify the optimal Transformer layer ALOPE systematically

[4] https://github.com/center-for-humans-and-machines

explores six candidate Transformer layers across all evaluated LLMs: the final layer (-1), intermediate layers (-7 and -11), a mid-range layer (-16), and lower-level layers (-20 and -24).

In addition to the layer-specific regression head approach (vanilla ALOPE), we extend the ALOPE framework with two additional strategies—dynamic weighting and multi-head regression, both of which leverage multiple Transformer layers to enhance translation quality estimation through multiple layer-level representation modelling.

3.3.1 Dynamic weighting

In the dynamic weighting approach, we extract sequence-level contextual representations from selected Transformer layers using auxiliary heads. Auxiliary heads are solely utilized to extract embeddings from specific Transformer layers and are not subjected to direct supervised training. To obtain a sequence-level embedding from the Transformer layer k, we select the final token's hidden state h_k. Instead of treating these selected layers equally, we assign a trainable scalar weight w_k to each layer k, and normalize these weights using the softmax function to identify each layer's contribution:

$$\alpha_k = \frac{\exp(w_k)}{\sum_{j=1}^{L} \exp(w_j)} \quad \text{for } k = 1, \ldots, L$$

These weights are initialized uniformly and subsequently optimized during training via backpropagation. The final embedding is a weighted sum of the individual layer embeddings:

$$\mathbf{h}_{\text{combined}} = \sum_{k=1}^{L} \alpha_k h_k$$

This representation is passed through a regression head to produce the predicted translation quality score. The model is optimized by minimizing the mean squared error between the predicted score and the ground truth. We refer to this method as *dynamic weighting* because it allows the model to dynamically learn to prioritize information from the most informative layers for the task of translation quality estimation.

3.3.2 Multi-head regression

In this approach, multiple regression heads are integrated into different Transformer layers, each tasked with independently approximating translation quality. To ensure a balanced optimization, a weighted aggregation of individual regression head losses is employed before back-propagation. The computed final loss is back-propagated through the model and updates the selected heads. This approach enables simultaneous learning across multiple layers. By optimizing a combined loss across selected layers, the framework effectively integrates multi-layered contextual information. During inference, the final translation quality prediction is derived by averaging the aggregated outputs from these regression heads.

3.4 Evaluation & Metrics

We use Spearman's correlation (Sedgwick, 2014) between the DA mean (averaged human-annotated DA scores from three or more annotators) and predictions as our evaluation metric. Additionally, we performed the Williams significance test (Graham & Baldwin, 2014; Williams, 1959) to assess whether the top-performing model for each language pair achieved a significantly higher Spearman correlation than other models.

4 Results and Discussion

Table 1 reports the Spearman correlation scores obtained under zero-shot evaluation, alongside the results from the ALOPE framework with regression heads placed at various Transformer layers (See section 3.3). The table also highlights cases where ALOPE yields improvements over standard instruction fine-tuning (SIFT) results with LLMs. Notably, the

	Model	En-Gu	En-Hi	En-Mr	En-Ta	En-Te	Et-En	Ne-En	Si-En	*Avg*
Zero-shot	LLaMA2-7B	0.006	-0.002	0.053	0.067	-0.016	0.168	0.153	0.144	0.072
	LLaMA3.1-8B	0.164	0.194	0.245	0.220	0.132	0.503	0.262	0.267	*0.248*
	LLaMA3.2-3B	0.095	0.095	0.116	0.128	0.098	0.359	0.216	0.120	*0.153*
	Aya-expanse-8B	0.123	0.135	0.089	0.117	0.049	0.315	0.352	0.330	*0.189*
	Avg	*0.097*	*0.106*	*0.126*	*0.133*	*0.066*	*0.336*	*0.246*	*0.215*	
TL (-1)	LLaMA2-7B	0.563 ↑	0.414 ↑	*0.609 ↑	0.525 ↑	0.356 ↑	*0.742 ↑	0.596 ↑	*0.565 ↑	0.546
	LLaMA3.1-8B	*0.594 ↑	*0.469 ↑	*0.620 ↑	0.567	0.363 ↑	0.734 ↑	0.647 ↑	0.547	*0.567*
	LLaMA3.2-3B	*0.604 ↑	*0.477 ↑	**0.636** ↑	0.580 ↑	0.348 ↑	0.735 ↑	*0.674 ↑	0.543 ↑	*0.575*
	Aya-expanse-8B	0.068	0.178	0.219	-0.006	0.275	0.115	0.012	0.077	*0.117*
	Avg	*0.457*	*0.385*	*0.521*	*0.416*	*0.335*	*0.581*	*0.482*	*0.433*	
TL (-7)	LLaMA2-7B	0.567 ↑	0.336 ↑	0.542 ↑	0.484 ↑	0.317 ↑	*0.739 ↑	0.606 ↑	**0.573**↑	0.520
	LLaMA3.1-8B	*0.590 ↑	*0.477 ↑	*0.625 ↑	0.528	**0.388** ↑	*0.744 ↑	0.638 ↑	0.544	*0.567*
	LLaMA3.2-3B	**0.606** ↑	**0.479**↑	*0.617↑	0.585↑	*0.369↑	**0.751**↑	0.664↑	*0.553↑	*0.578*
	Aya-expanse-8B	0.538 ↑	0.447 ↑	0.597 ↑	0.528 ↑	0.347 ↑	*0.741 ↑	0.646 ↑	0.544 ↑	*0.549*
	Avg	*0.575*†	*0.435*†	*0.595*†	*0.531*†	*0.355*†	*0.744*†	*0.639*†	*0.554*†	
TL (-11)	LLaMA2-7B	0.360	0.301	0.361	0.254	0.293↑	0.405	0.164	0.049	0.273
	LLaMA3.1-8B	0.514	0.412↑	*0.609↑	0.438	0.304↑	0.148	0.554	0.493	*0.434*
	LLaMA3.2-3B	*0.594↑	*0.476↑	0.605↑	**0.610**↑	*0.373↑	*0.748↑	*0.678↑	*0.560↑	*0.581*
	Aya-expanse-8B	0.490	0.411↑	0.572↑	0.445	0.336↑	0.569	0.453	0.439↑	*0.464*
	Avg	*0.489*	*0.400*	*0.537*	*0.437*	*0.327*	*0.467*	*0.462*	*0.385*	
TL (-16)	LLaMA2-7B	0.540 ↑	0.381 ↑	0.585 ↑	0.482 ↑	0.308 ↑	*0.751 ↑	0.580 ↑	*0.569 ↑	0.524
	LLaMA3.1-8B	0.558 ↑	0.453 ↑	0.602 ↑	0.523	0.350 ↑	*0.737 ↑	0.652 ↑	0.513	*0.548*
	LLaMA3.2-3B	0.557 ↑	*0.459 ↑	0.597 ↑	0.547	0.338 ↑	*0.745 ↑	**0.682**↑	*0.567 ↑	*0.561*
	Aya-expanse-8B	0.467	0.390 ↑	0.557 ↑	0.481	0.314 ↑	0.727 ↑	0.576 ↑	0.540 ↑	*0.506*
	Avg	*0.530*	*0.421*	*0.585*	*0.508*	*0.327*	*0.740*	*0.622*	*0.547*	
TL (-20)	LLaMA2-7B	0.470↑	0.405↑	0.544↑	0.460↑	0.338↑	0.684↑	0.508↑	0.534↑	0.493
	LLaMA3.1-8B	0.484	0.394↑	0.553↑	0.321	0.172	0.649	0.524	0.494	*0.449*
	LLaMA3.2-3B	0.430	0.408↑	0.579↑	0.303	0.286↑	0.601	0.488	0.464↑	*0.445*
	Aya-expanse-8B	0.437	0.300	0.488	0.263	0.287	0.483	0.438	0.395↑	*0.386*
	Avg	*0.455*	*0.377*	*0.541*	*0.337*	*0.271*	*0.604*	*0.490*	*0.472*	
TL (-24)	LLaMA2-7B	0.500↑	0.421↑	0.538↑	0.379	0.239	0.630↑	0.507↑	0.472↑	0.461
	LLaMA3.1-8B	0.421	0.378↑	0.552↑	0.330	0.290↑	0.515	0.530	0.464	*0.435*
	LLaMA3.2-3B	0.443	0.376	0.507	0.367	0.299↑	0.559	0.528	0.487↑	*0.446*
	Aya-expanse-8B	0.375	0.319	0.440	0.337	0.220	0.393	0.407	0.345	*0.354*
	Avg	*0.435*	*0.373*	*0.509*	*0.353*	*0.262*	*0.524*	*0.493*	*0.442*	

Table 1: Spearman correlation scores for zero-shot inference and ALOPE experiments across different Transformer layers (TL). TL(-1) refers to the final Transformer layer, TL(-7), TL(-11) are intermediate layers, TL(-16) is mid-layer, and TL(-20), TL(-24) are lower-level layers. ↑ indicates cases where ALOPE outperforms standard instruction fine-tuning (SIFT) for a given language pair (Correlation scores for SIFT can be found in Appendix B). Bolded values denote the highest correlation scores per language pair; (†) marks the highest average across Transformer layers; (*) indicates the statistically insignificant Spearman scores compared to the best scores. The dashed line separates language pairs where English is the target language.

performance under zero-shot settings is substantially lower across all eight low-resource language pairs when compared to both SIFT and ALOPE. When benchmarked against the best Spearman scores from SIFT (refer to Appendix B), ALOPE obtains the best correlation scores for all evaluated language pairs.

4.1 Layer-specific adaptation for QE

The experimental results obtained using the ALOPE framework, presented in Table 1, reveal that the most effective performance across language pairs is not achieved when the regression head is placed at the final Transformer layer (TL-1), but rather when it is positioned at intermediate layers (TL-7, TL-11) between the final and mid-level layers.

Among all layers evaluated, TL-7 consistently yielded the highest Spearman correlation scores for the majority of language pairs, demonstrating both top-performing outcomes and a high number of instances with statistically insignificant differences from the best scores. The layer-wise averages shown in Table 1 also confirm that TL-7 consistently offers the highest average correlation across models and language pairs. These suggest that TL-7 provides robust and reliable contextual representations suitable for quality estimation of low-resource language pairs.

Additionally, TL-11, while producing the highest score for En-Ta, exhibited statistically insignificant differences from the best-performing layer across the other seven language pairs. This further reinforces the effectiveness of intermediate layers in capturing cross-lingual alignment. Collectively, these results indicate that optimal layer selection is a key factor in enhancing QE performance with LLMs and that intermediate Transformer layers are better suited for extracting representations that support accurate quality prediction in low-resource machine translation tasks.

At the mid-layer (TL-16), the Ne-En language pair achieved the highest correlation, while three other language pairs (En-Hi, Et-En, and Si-En) exhibited performance that was statistically indistinguishable from the best scores, as shown in Table 1. Notably, three out of these four language pairs share a common characteristic: English serves as the target language in the translation direction. This pattern suggests that cross-lingual transfer learning in Transformer-based LLMs reaches representational consistency earlier from the mid-layer onwards, when English is the target language. The earlier stabilization may be attributed to the extensive exposure of LLMs to English during pre-training, which enhances cross-lingual alignment when generating English translations (Kargaran et al., 2024). Furthermore, as shown in Table 1, models consistently achieve higher average correlation scores when English is the target language in MT, indicating stronger cross-lingual transfer in that direction. Notably, Et–En yields the highest layer-wise average across all experimental settings, likely due to the shared Latin-based script, reinforcing prior findings on the role of script similarity in enhancing cross-lingual transfer (Sindhujan et al., 2025).

The overall results from ALOPE indicate a decrease in correlation scores beyond the mid-Transformer layer (TL-16) for most models and language pairs. This aligns with the observation of Kargaran et al. (2024), that as embeddings for low-resource languages when progressed through the mid-Transformer layers, they became more aligned with the primary language of the LLMs. This improved alignment likely contributes to better cross-lingual representation at intermediate layers (TL-11, TL-7) after the mid-layers, and may also explain the decreased performance observed in the lower-level Transformer layers (TL-20, and TL-24), where such alignment has not yet stabilized.

4.1.1 Model-wise analysis

The best-performing models under the ALOPE framework consistently belonged to the LLaMA family. Notably, the LLaMA 3.2 model yielded the highest Spearman correlation scores for six out of eight language pairs: En to {Gu, Hi, Mr, Ta} and {Et, Ne} to En. LLaMA 3.1 demonstrated the best performance for En–Te, while LLaMA 2–7B outperformed other models on Si–En. These results highlight the effectiveness of LLaMA-based models for cross-lingual tasks over other open-source models.

Furthermore, statistical analysis using the Williams significance test revealed that although LLaMA 3.2 did not yield the absolute highest correlation scores for En–Te or Si–En, its performance at TL–7 and TL–11 was statistically indistinguishable from the best-performing models (as detailed in Table 1). Notably, LLaMA 3.2 attained this level of performance despite having the smallest parameter size (3 billion) among all the models evaluated in this study. This suggests that model size alone is not the sole determinant of the performance of a cross-lingual task such as QE and highlights the efficiency and adaptability of LLaMA 3.2 within resource-constrained environments. The Aya model also demonstrated correlation scores comparable to LLaMA models, with one notable exception, when the ALOPE experimented with the final Transformer layer (TL-1), its performance exhibited a significant decrease. The possible reason for this could be that the final layer of Aya is

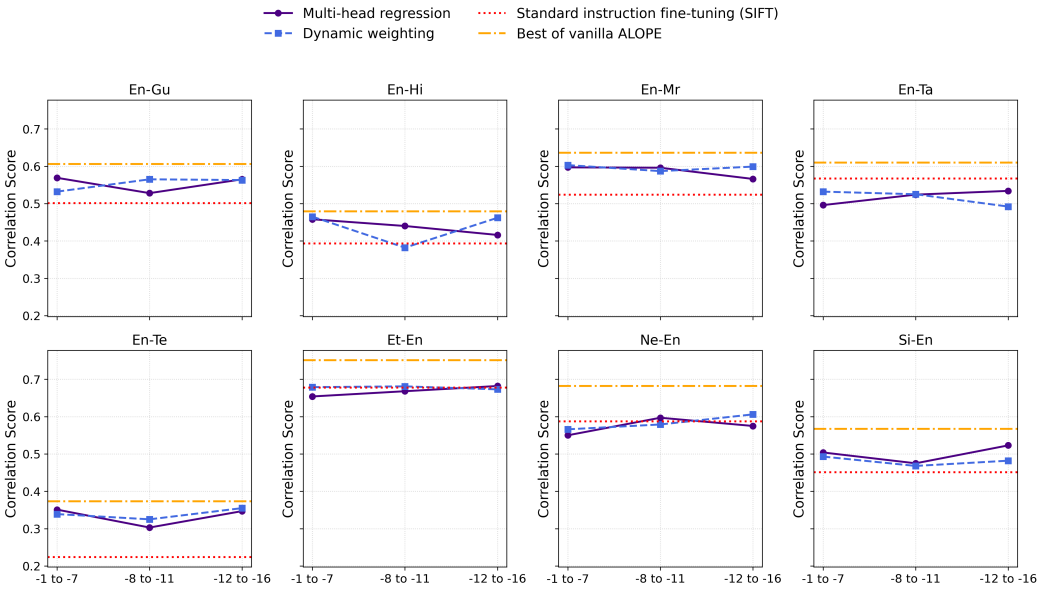

Figure 3: LLaMA3.2-3B results for dynamic-weighting and multi-head regression compared against standard instruction fine-tuning results and best of Vanilla ALOPE. Y-axis represents the correlation score and the X-axis represents the selected range of Transformer layers for dynamic weighting and multi-head regression experiments.

less aligned with quality estimation signals, whereas its intermediate representations still capture useful features. Overall, all the LLMs demonstrated improved performance when using ALOPE compared to SIFT, when the regression heads were attached to the optimal Transformer layers.

4.2 Dynamic weighting & Multi-head regression for QE

To further evaluate the benefits of leveraging multiple Transformer layers, dynamic weighting and multi-head regression experiments were conducted using the LLaMA 3.2-3B model, which had previously demonstrated the most consistent performance across language pairs (Section 4.1). The experiments focused on three Transformer layer groupings: TL-1 to TL-7, TL-8 to TL-11, and TL-12 to TL-16. As illustrated in Figure 3, both approaches yielded improvements over SIFT for En-{Gu, Mr, Te} and Si–En. For the En-Hi language pair, the multi-head regression approach outperformed SIFT across all three layer groupings, whereas dynamic weighting exhibited a slight performance drop in the TL-8 to TL-11 configuration. In the case of Et-En and Ne-En, both multi-layer strategies delivered comparable performance to SIFT, with only marginal deviations observed. However, the En-Ta pair continued to yield suboptimal results under both approaches, resembling its weaker performance trends observed in previous experiments. This underscores the unique challenges associated with En-Ta in cross-lingual QE tasks.

In Figure 3, Vanilla ALOPE represents the best correlation scores obtained through layer-specific adaptation with LLaMA3.2-3B model, as detailed in Section 4.1. When comparing the performance of multi-layer approaches with Vanilla ALOPE's layer-specific adaptation, Vanilla ALOPE generally achieves higher correlation scores. Nevertheless, both multi-layer approaches demonstrate improvements over SIFT for the majority of the language pairs. These results highlight that multi-layer methods can also effectively leverage complementary information from multiple Transformer layers, making them valuable alternatives for efficient LLM-based translation quality estimation.

4.3 Comparative analysis: ALOPE *vs.* QE frameworks & Monolingual setting

The ALOPE approach demonstrated consistent improvements in correlation scores over SIFT across all language pairs and achieved results comparable to those of established

Method	En-Gu	En-Hi	En-Mr	En-Ta	En-Te	Et-En	Ne-En	Si-En
ALOPE*	0.606	0.479	**0.636**	0.610	**0.388**	0.751	0.682	0.573
SIFT-LLMs*	0.555	0.393	0.530	0.586	0.290	0.728	0.618	0.558
TransQuest*	0.630	0.478	0.606	0.603	0.358	0.760	0.718	0.579
CometKiwi	**0.637**	**0.546**	0.635	**0.616**	0.338	**0.860**	**0.789**	**0.703**

Table 2: Comparison of highest Spearman correlation scores across ALOPE, standard instruction fine-tuning (SIFT) with LLMs, and encoder-based QE baselines—TransQuest and CometKiwi. Methods marked with (*) are trained using the datasets described in Appendix A. The CometKiwi results are from the publicly available model pre-trained on over 940K QE data samples (Rei et al., 2023).

pre-trained encoder-based QE frameworks, as shown in Table 2. Notably, for the En-Mr and En-Te language pairs, ALOPE even surpassed the SOTA performance of these frameworks. While pre-trained encoder-based models such as TransQuest (InfoXLM-Large) and CometKiwi (XLM-R-XL) continue to show strong performance, they require dedicated training pipelines and a significant amount of data for fine-tuning. In contrast, ALOPE offers a flexible and modular solution that can be applied to pre-deployed LLMs without extensive reconfiguration, making it a practical solution for performing QE.

Despite being based on LLMs, ALOPE demonstrates favourable memory efficiency when compared to encoder-based QE frameworks. Although LLMs are often considered more resource-intensive, our implementation of ALOPE—augmented with LoRA-adapted lightweight regression heads exhibits competitive GPU memory usage, as detailed in Appendix C. This indicates that ALOPE can be deployed with overheads comparable to, or even lower than, those of established encoder-based models, highlighting its practicality in resource-constrained environments. Finally, it is also scalable to a unified evaluation and correction framework where an additional causal head can provide translation error reasoning-based corrections or perform automatic post-editing.

To evaluate the generalizability of ALOPE for regression tasks beyond cross-lingual quality estimation, we conducted an ablation study, which is detailed in the Appendix E, observing its performance on a monolingual regression task. Results demonstrated that ALOPE consistently outperforms standard instruction fine-tuning and baseline scores in the monolingual setting, achieving the highest scores across evaluated languages- English, Spanish and Arabic. These findings highlight ALOPE's ability to effectively identify optimal transformer layers for enhanced performance in both monolingual and cross-lingual settings.

5 Conclusion

This paper introduced ALOPE, a novel framework for improving segment-level translation quality estimation using LLMs. By integrating regression heads with LoRA and strategically leveraging intermediate Transformer layers, ALOPE outperforms standard instruction fine-tuning results of LLMs and achieves results on par with, or better than, encoder-based QE frameworks. Our experiments demonstrate that intermediate Transformer layers, particularly TL-7, yield more effective cross-lingual representations for QE, while performance tends to decrease in layers beyond the mid-range. Notably, LLaMA 3.2 emerged as the most effective model across language pairs, despite having the smallest parameter size, underscoring that parameter count is not the sole indicator of QE effectiveness. Furthermore, early stabilization of the performance was observed when English served as the target language in translation, especially around mid-layer depths. Additionally, the framework introduces dynamic weighting and multi-head regression approaches, which leverage a selective set of multiple Transformer layers to perform QE. ALOPE demonstrates competitive GPU efficiency, reinforcing its practicality as a flexible and scalable solution for deploying LLMs in resource-constrained settings to perform QE. We also envision this framework to be extended for reasoning over MT errors informed by a regression-head predicted DA, providing more reliability to error reasoning and MT assessment.

References

Yujin Baek, Zae Myung Kim, Jihyung Moon, Hyunjoong Kim, and Eunjeong Park. Patquest: Papago translation quality estimation. In *Proceedings of the Fifth Conference on Machine Translation*, pp. 991–998, 2020.

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. Findings of the WMT 2023 shared task on quality estimation. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz (eds.), *Proceedings of the Eighth Conference on Machine Translation*, pp. 629–653, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.52. URL https://aclanthology.org/2023.wmt-1.52.

Rochelle Choenni, Dan Garrette, and Ekaterina Shutova. Cross-lingual transfer with language-specific subnetworks for low-resource dependency parsing. *Computational Linguistics*, pp. 613–641, September 2023. doi: 10.1162/coli_a_00482. URL https://aclanthology.org/2023.cl-3.3.

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

Yvette Graham and Timothy Baldwin. Testing for significance of increased correlation with human judgment. In Alessandro Moschitti, Bo Pang, and Walter Daelemans (eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 172–176, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1020. URL https://aclanthology.org/D14-1020.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. Continuous measurement scales in human evaluation of machine translation. In Antonio Pareja-Lora, Maria Liakata, and Stefanie Dipper (eds.), *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pp. 33–41, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://aclanthology.org/W13-2305.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Julia Ive, Frédéric Blain, and Lucia Specia. Deepquest: a framework for neural-based quality estimation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 3146–3157, 2018.

Amir Hossein Kargaran, Ali Modarressi, Nafiseh Nikeghbal, Jana Diesner, François Yvon, and Hinrich Schütze. Mexa: Multilingual evaluation of english-centric llms via cross-lingual alignment, 2024. URL https://arxiv.org/abs/2410.05873.

Amir Hossein Kargaran, Ali Modarressi, Nafiseh Nikeghbal, Jana Diesner, François Yvon, and Hinrich Schuetze. MEXA: Multilingual evaluation of English-centric LLMs via cross-lingual alignment. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 27001–27023, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. URL https://aclanthology.org/2025.findings-acl.1385/.

Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André FT Martins. Openkiwi: An open source framework for quality estimation. *arXiv preprint arXiv:1902.08646*, 2019.

Hyun Kim, Hun-Young Jung, Hongseok Kwon, Jong-Hyeok Lee, and Seung-Hoon Na. 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.

Tom Kocmi and Christian Federmann. Large language models are state-of-the-art evaluators of translation quality. In Mary Nurminen, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, Eva Vanmassenhove, Sergi Alvarez Vidal, Nora Aranberri, Mara Nunziatini, Carla Parra Escartín, Mikel Forcada, Maja Popovic, Carolina Scarton, and Helena Moniz (eds.), *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pp. 193–203, Tampere, Finland, June 2023. European Association for Machine Translation. URL https://aclanthology.org/2023.eamt-1.19.

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

Arle Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK, November 28-29 2013. Aslib. URL https://aclanthology.org/2013.tc-1.6/.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. SemEval-2018 task 1: Affect in tweets. In Marianna Apidianaki, Saif M. Mohammad, Jonathan May, Ekaterina Shutova, Steven Bethard, and Marine Carpuat (eds.), *Proceedings of the 12th International Workshop on Semantic Evaluation*, pp. 1–17, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-1001. URL https://aclanthology.org/S18-1001.

Joao Moura, Miguel Vera, Daan van Stigt, Fabio Kepler, and André FT Martins. Ist-unbabel participation in the wmt20 quality estimation shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pp. 1029–1036, 2020.

Vandan Mujadia, Pruthwik Mishra, Arafat Ahsan, and Dipti M. Sharma. Towards large language model driven reference-less translation evaluation for English and Indian language. In Jyoti D. Pawar and Sobha Lalitha Devi (eds.), *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pp. 357–369, Goa University, Goa, India, December 2023. NLP Association of India (NLPAI). URL https://aclanthology.org/2023.icon-1.28.

Xuan-Phi Nguyen, Mahani Aljunied, Shafiq Joty, and Lidong Bing. Democratizing LLMs for low-resource languages by leveraging their English dominant abilities with linguistically-diverse prompts. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3501–3516, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.192.

Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Niccolò Campolungo, and Roberto Navigli. Matese: Machine translation evaluation as a sequence tagging problem. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pp. 569–577, 2022.

Trinh Pham, Khoi Le, and Anh Tuan Luu. UniBridge: A unified approach to cross-lingual transfer learning for low-resource languages. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3168–3184, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.174. URL https://aclanthology.org/2024.acl-long.174.

Shenbin Qian, Archchana Sindhujan, Minnie Kabra, Diptesh Kanojia, Constantin Orasan, Tharindu Ranasinghe, and Fred Blain. What do large language models need for machine translation evaluation? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 3660–3674, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.214. URL https://aclanthology.org/2024.emnlp-main.214/.

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. TransQuest: Translation quality estimation with cross-lingual transformers. In Donia Scott, Nuria Bel, and Chengqing Zong (eds.), *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 5070–5081, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.445. URL https://aclanthology.org/2020.coling-main.445/.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri (eds.), *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pp. 634–645, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.60/.

Ricardo Rei, Nuno M. Guerreiro, JosÃ© Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz (eds.), *Proceedings of the Eighth Conference on Machine Translation*, pp. 841–848, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.73. URL https://aclanthology.org/2023.wmt-1.73/.

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

Archchana Sindhujan, Diptesh Kanojia, and Constantin Orasan. Optimizing quality estimation for low-resource language translations: Exploring the role of language relatedness. In *Proceedings of the Conference New Trends in Translation and Technology 2024*. Incoma, 2024.

Archchana Sindhujan, Diptesh Kanojia, Constantin Orasan, and Shenbin Qian. When LLMs struggle: Reference-less translation evaluation for low-resource languages. In Hansi Hettiarachchi, Tharindu Ranasinghe, Paul Rayson, Ruslan Mitkov, Mohamed Gaber, Damith Premasiri, Fiona Anting Tan, and Lasitha Uyangodage (eds.), *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pp. 437–459, Abu Dhabi, United Arab Emirates, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.loreslm-1.33/.

Lucia Specia, Kashif Shah, José GC De Souza, and Trevor Cohn. 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.

Lucia Specia, Gustavo Henrique Paetzold, and Carolina Scarton. Multi-level Translation Quality Prediction with QUEST++. pp. 115–120. Association for Computational Linguistics and The Asian Federation of Natural Language Processing, 7 2015. doi: 10.3115/v1/P15-4020. URL https://aclanthology.org/P15-4020.

Tu Vu, Kalpesh Krishna, Salaheddin Alzubi, Chris Tar, Manaal Faruqui, and Yun-Hsuan Sung. Foundational autoraters: Taming large language models for better automatic evaluation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17086–17105, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.949. URL https://aclanthology.org/2024.emnlp-main.949/.

E.J. Williams. *Regression Analysis*. WILEY SERIES in PROBABILITY and STATISTICS: APPLIED PROBABILITY and STATIST ICS SECTION Series. Wiley, 1959. ISBN 9780471946779. URL https://books.google.co.uk/books?id=HXdqAAAAMAAJ.

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. Findings of the WMT 2022 shared task on quality estimation. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri (eds.), *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pp. 69–99, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.3.

Chrysoula Zerva, Frederic Blain, José G. C. De Souza, Diptesh Kanojia, Sourabh Deoghare, Nuno M. Guerreiro, Giuseppe Attanasio, Ricardo Rei, Constantin Orasan, Matteo Negri, Marco Turchi, Rajen Chatterjee, Pushpak Bhattacharyya, Markus Freitag, and André Martins. Findings of the quality estimation shared task at WMT 2024: Are LLMs closing the gap in QE? In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz (eds.), *Proceedings of the Ninth Conference on Machine Translation*, pp. 82–109, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.wmt-1.3. URL https://aclanthology.org/2024.wmt-1.3/.

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. A survey of large language models, 2023.

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

Shaolin Zhu, Supryadi, Shaoyang Xu, Haoran Sun, Leiyu Pan, Menglong Cui, Jiangcun Du, Renren Jin, António Branco, and Deyi Xiong. Multilingual large language models: A systematic survey, 2024. URL https://arxiv.org/abs/2411.11072.

A Dataset

	En-Gu	En-Hi	En-Mr	En-Ta	En-Te	Et-En	Ne-En	Si-En
Train	7000	7000	26000	7000	7000	7000	7000	7000
Test	1000	1000	699	1000	1000	1000	1000	1000

Table 3: Dataset splits of QE dataset utilized in our study. Experiments were conducted with 8 low-resource language pairs to evaluate model performance.

B Results of standard instruction fine-tuned LLMs (SIFT)

Models	**En-Gu**	**En-Hi**	**En-Mr**	**En-Ta**	**En-Te**	**Et-En**	**Ne-En**	**Si-En**	***Avg***
LLaMA2-7B	0.454	0.319	0.466	0.449	0.266	0.591	0.478	0.374	0.425
LLaMA3.1-8B	0.555	0.368	**0.530**	0.586	0.233	**0.728**	**0.618**	**0.558**	0.522
LLaMA3.2-3B	0.501	**0.393**	0.524	**0.567**	0.224	0.678	0.587	0.451	0.491
Aya-expanse-8B	**0.528**	0.388	0.515	0.496	**0.290**	0.605	0.568	0.384	0.472
Avg	0.510	0.367	0.509	0.525	0.253	0.650	0.563	0.442	

Table 4: Spearman correlation score obtained with standard instruction fine-tuning. Bolded values denote the highest Spearman score obtained for each language pair

C GPU memory utilization

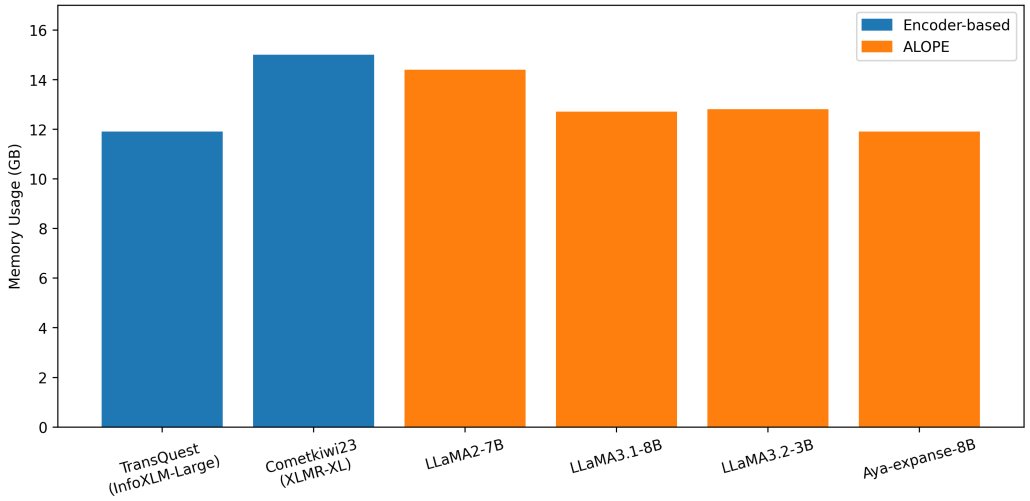

Figure 4: GPU memory utilization comparison between encoder-based QE models and various LLM-based ALOPE configurations.

D Language pair-wise performance across different Transformer layers

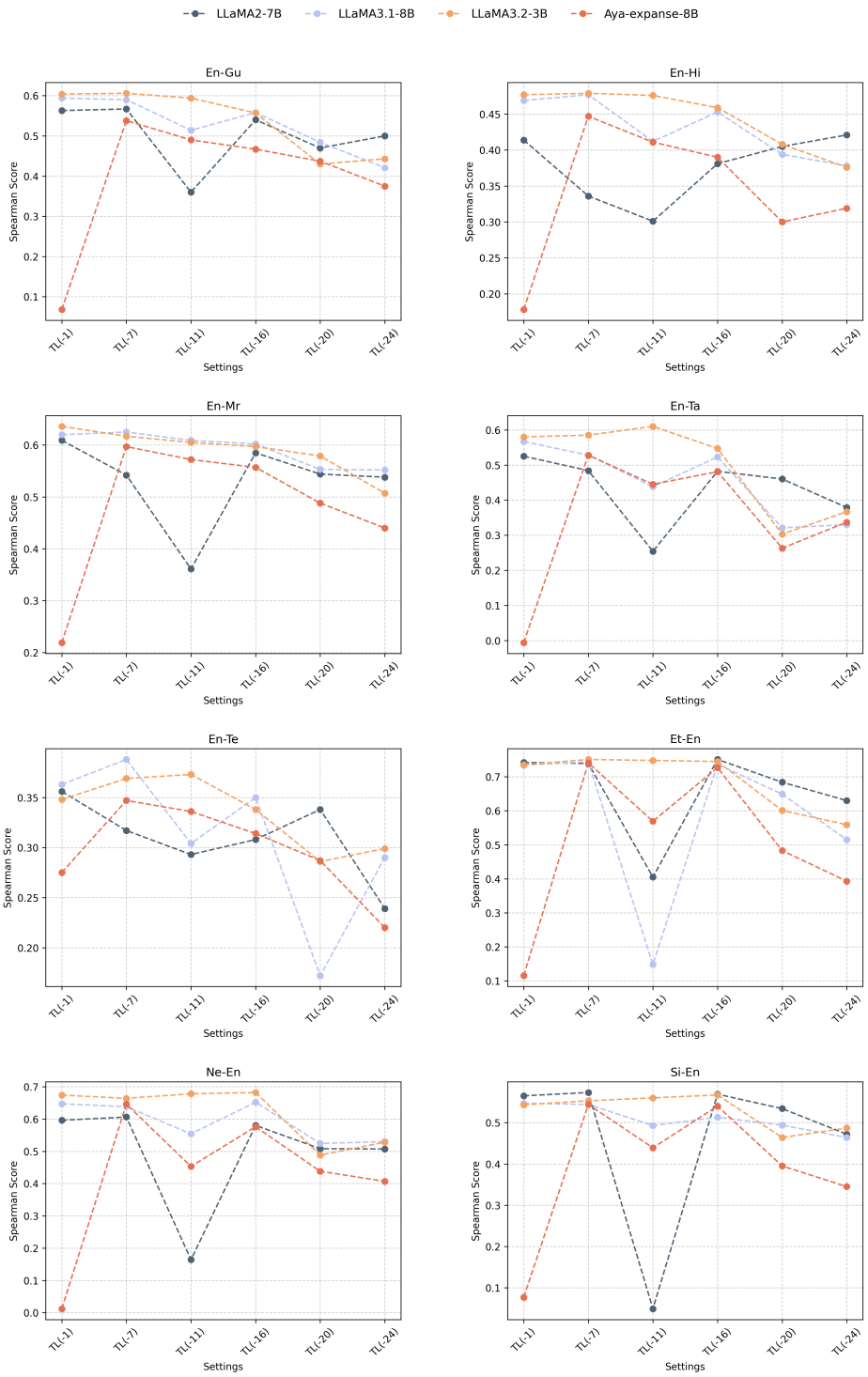

Figure 5: Spearman correlation scores obtained across different Transformer layers with ALOPE layer-specific regression experiments.

E Assessing the generalization of ALOPE across regression tasks: An ablation study

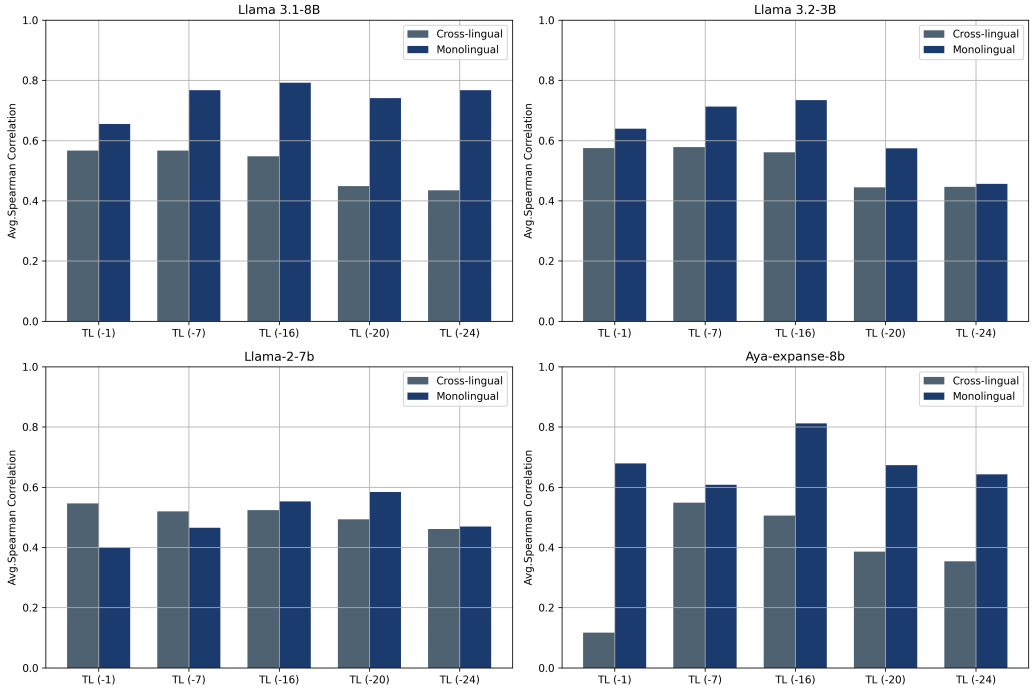

Figure 6: Model-wise average Spearman correlation scores obtained with layer-specific ALOPE approach, comparing performance on monolingual versus cross-lingual regression tasks.

The ALOPE framework consistently outperformed standard instruction fine-tuned LLMs in the cross-lingual regression task of quality estimation across all language pairs and delivered results comparable to leading encoder-based QE models (Table 2). For En-Mr and En-Te, it even exceeded state-of-the-art performance. To better understand whether the performance gains observed with ALOPE are specific to cross-lingual QE or indicative of broader regression capabilities, we conducted an ablation study using a monolingual regression task.

For the comparative analysis, we utilized the SemEval-2018 shared task dataset (Mohammad et al., 2018), which focuses on predicting emotion intensity scores, real-valued outputs ranging from 0 to 1 based on tweet content. The dataset spans four emotion categories and comprises monolingual data in English, Spanish, and Arabic. SemEval was selected for its structural similarity to the QE dataset, as both tasks involve predicting a continuous score based on the text input. However, the key distinction lies in language composition: SemEval is monolingual, whereas QE contains source-target translation data, which is cross-lingual. This contrast makes SemEval a suitable benchmark for evaluating ALOPE's performance in monolingual versus cross-lingual regression settings.

Experiments were conducted using the same set of LLMs within the ALOPE framework, demonstrating its flexibility in adapting to monolingual regression tasks. As shown in the Table 6, all three languages obtained the best score with ALOPE for emotion intensity score prediction showing improved correlation score over the baseline, *i.e.*, best score obtained for the SemEval-2018 shared task (Mohammad et al., 2018) and standard instruction fine-tuned results of LLMs. This shows the effectiveness of ALOPE for LLM-based regression tasks.

Although direct comparison between monolingual and cross-lingual tasks is not feasible due to their differing objectives, we analyzed model-wise average Spearman correlation scores across selected Transformer layers (see Figure 6). This graph shows that the monolingual

You are provided with a tweet and a specific emotion. Your task is to determine the intensity of that emotion as expressed in the tweet. The intensity score should be a real value between 0 (indicating the least amount of emotion) and 1 (indicating the most intense level of that emotion). Focus on understanding the context, tone, and emotion of the tweet to accurately determine the emotional intensity.

Tweet: {tweet}

Emotion: {emotion}

Provide an intensity score between 0 and 1 that best represents the emotional state of the person tweeting.

Figure 7: Prompt for emotion intensity prediction

	English	**Spanish**	**Arabic**
Train	7102	4544	3376
Test	4068	2616	1563

Table 5: Data count of each split utilized for emotion intensity score prediction.

regression task consistently achieves higher average correlation scores across most of the Transformer layers, including the lower-level layers (TL:-20 and -24). In contrast, cross-lingual QE tasks show improved performance predominantly in the intermediate layers beyond the mid-level. These findings reinforce ALOPE's effectiveness in identifying and leveraging the most suitable Transformer layers for cross-lingual quality estimation, thereby enhancing performance in low-resource settings.

Building on this analysis, the ALOPE framework demonstrates strong adaptability and consistently outperforms standard LLM baselines across both monolingual and cross-lingual regression tasks, highlighting its generalizability and effectiveness in identifying optimal transformer layers for varied language settings.

	Model	English		Spanish		Arabic	
		ρ	r	ρ	r	ρ	r
SIFT	llama-2-7b	0.766	0.767	0.775	0.778	0.481	0.486
	llama 3.1-8B	0.794	0.792	0.797	0.798	0.662	0.670
	llama 3.2-3B	0.777	0.775	0.755	0.752	0.546	0.558
	aya-expanse-8b	0.779	0.782	0.807	0.807	0.721	0.727
	Avg	*0.779*	-	*0.783*	-	*0.602*	-
TL (-1)	llama-2-7b	0.755	0.759	0.321	0.320	0.123	0.116
	llama 3.1-8B	0.786	0.788	0.709	0.709	0.471	0.480
	llama 3.2-3B	0.773	0.771	0.706	0.705	0.438	0.440
	aya-expanse-8b	0.767	0.771	0.760	0.760	0.510	0.516
	Avg	*0.770*	-	*0.624*	-	*0.385*	-
TL (-7)	llama-2-7b	0.747	0.748	0.410	0.407	0.239	0.238
	llama 3.1-8B	0.823	0.825	0.812	0.815	0.665	0.670
	llama 3.2-3B	0.801	0.803	0.787	0.791	0.553	0.556
	aya-expanse-8b	0.686	0.683	0.657	0.662	0.482	0.490
	Avg	*0.764*	-	*0.666*	-	*0.485*	-
TL (-16)	llama-2-7b	0.801	0.804	0.748	0.750	0.111	0.104
	llama 3.1-8B	0.830	0.829	**0.840**	0.825	0.707	0.712
	llama 3.2-3B	0.802	0.804	0.791	0.794	0.611	0.611
	aya-expanse-8b	0.826	0.830	0.839	0.841*	**0.769**	0.779*
	Avg	*0.814*	-	*0.804*	-	*0.549*	-
TL (-20)	llama-2-7b	0.821	0.823	0.753	0.756	0.177	0.186
	llama 3.1-8B	**0.831**	0.831*	0.833	0.835	0.559	0.567
	llama 3.2-3B	0.801	0.802	0.741	0.744	0.178	0.166
	aya-expanse-8b	0.784	0.788	0.757	0.758	0.480	0.491
	Avg	*0.809*	-	*0.771*	-	*0.348*	-
TL (-24)	llama-2-7b	0.745	0.742	0.381	0.375	0.281	0.288
	llama 3.1-8B	0.806	0.808	0.787	0.790	0.708	0.717
	llama 3.2-3B	0.635	0.632	0.311	0.307	0.420	0.420
	aya-expanse-8b	0.733	0.731	0.619	0.615	0.576	0.589
	Avg	*0.730*	-	*0.525*	-	*0.496*	-
	SemEval Task Best	-	0.799	-	0.738	-	0.685

Table 6: Spearman's (ρ) and Pearson's (r) correlation scores obtained with ALOPE framework for the monolingual regression task of emotion intensity prediction across multiple language pairs. The last row in the table shows the *best* Pearson correlation score obtained in SemEval-2018 for the emotion-intensity score prediction task (Mohammad et al., 2018). The bolded values show the highest Spearman score obtained and asterisks (*) indicate the highest Pearson score obtained.