ALOPE: Adaptive Layer Optimization for Translation Quality Estimation using LLMs

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

Surrey Institute for People-centred AI, Centre for Translation Studies, School of Computer Science and Electronic Engineering
{a.sindhujan, s.qian, c.orasan, d.kanojia}@surrey.ac.uk,matthewchancc@gmail.com

Introduction

Quality Estimation (QE) for Machine Translation is essential for evaluating translations without references, but it remains challenging for Large Language Models (LLMs), which are optimized for causal language modelling rather than fine-grained regression tasks. Existing LLM-based QE approaches often rely only on the final Transformer layer, ignoring potentially valuable intermediate representations that may provide better cross-lingual alignment, particularly important for low-resource languages. To address this, we introduce **ALOPE**.

What is ALOPE?

ALOPE is an LLM-based framework for estimating the quality of Machine Translated outputs. ALOPE restructures the Transformer representations through layer-wise adaptation with LoRA and regression task heads. It leverages intermediate Transformer layers that capture cross-lingual contextual features more effectively for regression-based QE prediction. Our framework shows improvements over various existing LLM-based QE approaches for low-resourced languages. Overall, ALOPE offers a scalable and efficient solution, bridging the gap between LLMs and state-of-the-art QE frameworks.

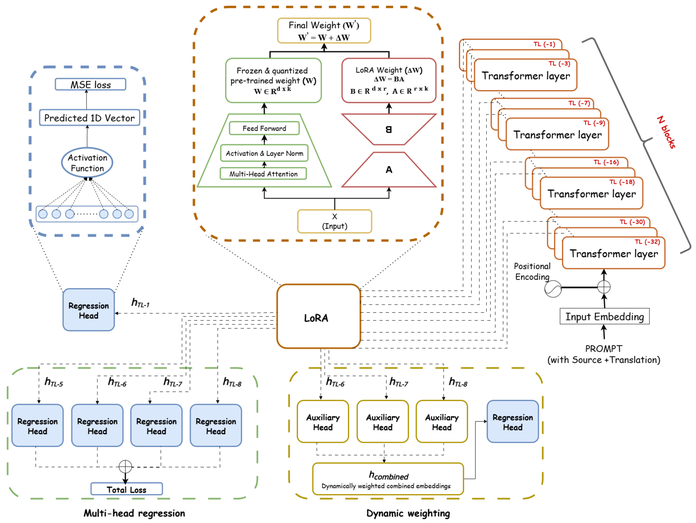
Experimental Settings

- **Models:** Experiments conducted with open-source LLMs (≤8B parameters) LLaMA2-7B, LLaMA3.1-8B, LLaMA3.2-3B, and Aya-expanse-8B
- **Prompt:** GEMBA prompt is utilized for all the experiments.
- **Language Pairs:** English to {Gujarathi, Hindi, Marathi, Tamil, Telugu} and {Estonian, Nepali and Sinhala} to English.
- **Training strategies:** Zero-shot inference, Standard Instruction Fine-Tuning (SIFT), and the proposed ALOPE framework based Instruction Tuning with regression heads + LoRA. Multilingual training was performed combining the eight low-resource language pairs and evaluation with language-specific test sets.
- **Evaluation:** Performance measured using Spearman's correlation between predicted and mean value of human-annotated DA scores.

ALOPE Approaches

- **Layer-specific:** Attaches regression heads to different Transformer layers, enabling the identification of the most effective layer representations for translation quality estimation.
- **Dynamic weighting:** Extracts embeddings from multiple Transformer layers and learns trainable weights, combining them into a single representation that prioritizes the most informative layers for QE.
- **Multi-head regression:** Integrates multiple regression heads across Transformer layers, jointly optimizes their losses during training, and averages their outputs at inference to capture multi-layer contextual information for QE.

ALOPE Framework

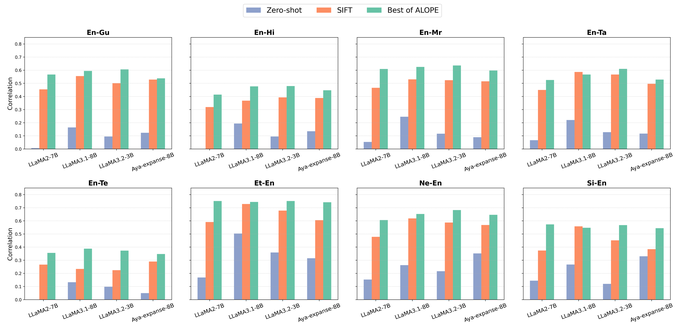


Transformer Layer-Specific Results with ALOPE

	Model	En-Gu	En-Hi	En-Mr	En-Ta	En-Te	Et-En	Ne-En	Si-En
TL (-1)	LLaMA2-7B	0.563 ↑	0.414 ↑	*0.609 ↑	0.525 ↑	0.356 ↑	*0.742 ↑	0.596 ↑	*0.565 ↑
	LLaMA3.1-8B	*0.594 ↑	*0.469 ↑	*0.620 ↑	0.567	0.363 ↑	0.734 ↑	0.647 ↑	0.547
	LLaMA3.2-3B	*0.604 ↑	*0.477 ↑	**0.636** ↑	0.580 ↑	0.348 ↑	0.735 ↑	*0.674 ↑	0.543 ↑
	Aya-expanse-8B	0.068	0.178	0.219	-0.006	0.275	0.115	0.012	0.077
TL (-7)	LLaMA2-7B	0.567 ↑	0.336 ↑	0.542 ↑	0.484 ↑	0.317 ↑	*0.739 ↑	0.606 ↑	**0.573**↑
	LLaMA3.1-8B	*0.590 ↑	*0.477 ↑	*0.625 ↑	0.528	**0.388** ↑	*0.744 ↑	0.638 ↑	0.544
	LLaMA3.2-3B	**0.606** ↑	**0.479**↑	*0.617↑	0.585↑	*0.369↑	**0.751**↑	0.664↑	*0.553↑
	Aya-expanse-8B	0.538 ↑	0.447 ↑	0.597 ↑	0.528 ↑	0.347 ↑	*0.741 ↑	0.646 ↑	0.544 ↑
TL (-11)	LLaMA2-7B	0.360	0.301	0.361	0.254	0.293↑	0.405	0.164	0.049
	LLaMA3.1-8B	0.514	0.412↑	*0.609↑	0.438	0.304↑	0.148	0.554	0.493
	LLaMA3.2-3B	*0.594↑	*0.476↑	0.605↑	**0.610**↑	*0.373↑	*0.748↑	*0.678↑	*0.560↑
	Aya-expanse-8B	0.490	0.411↑	0.572↑	0.445	0.336↑	0.569	0.453	0.439↑
TL (-16)	LLaMA2-7B	0.540 ↑	0.381 ↑	0.585 ↑	0.482 ↑	0.308 ↑	*0.751 ↑	0.580 ↑	*0.569 ↑
	LLaMA3.1-8B	0.558 ↑	0.453 ↑	0.602 ↑	0.523	0.350 ↑	*0.737 ↑	0.652 ↑	0.513
	LLaMA3.2-3B	0.557 ↑	*0.459 ↑	0.597 ↑	0.547	0.338 ↑	*0.745 ↑	**0.682**↑	*0.567 ↑
	Aya-expanse-8B	0.467	0.390 ↑	0.557 ↑	0.481	0.314 ↑	0.727 ↑	0.576 ↑	0.540 ↑

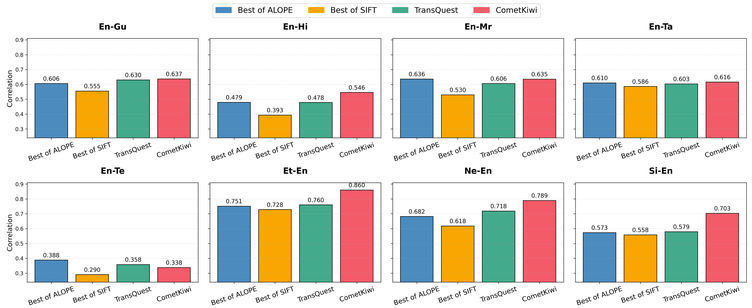
Zero-shot vs SIFT vs ALOPE

We compare results across Zero-shot, SIFT, and ALOPE for each model and language pair. In most cases, ALOPE achieves the highest values, and it **consistently yields the best correlation scores across all eight language pairs**.



Encoder vs Decoder based QE Approaches

ALOPE demonstrates consistent improvements over SIFT and performs on par with strong encoder-based QE frameworks. For the En–Mr and En–Te language pairs, it even surpasses the current state-of-the-art correlation score of CometKiwi.



Conclusion

We introduced ALOPE, a framework that integrates regression heads with LoRA to leverage intermediate Transformer layers for translation quality estimation. Experiments across eight low-resource language pairs show that ALOPE consistently outperforms standard instruction fine-tuning and achieves results comparable to strong encoder-based QE frameworks. ALOPE also demonstrates competitive GPU efficiency, reinforcing its practicality as a flexible and scalable solution for deploying LLMs in resource-constrained settings to perform QE.