Optimizing Large Language Models for Low-resource Quality Estimation

Diptesh Kanojia

12th Workshop for Asian Translation (WAT)
@
AACL 2025

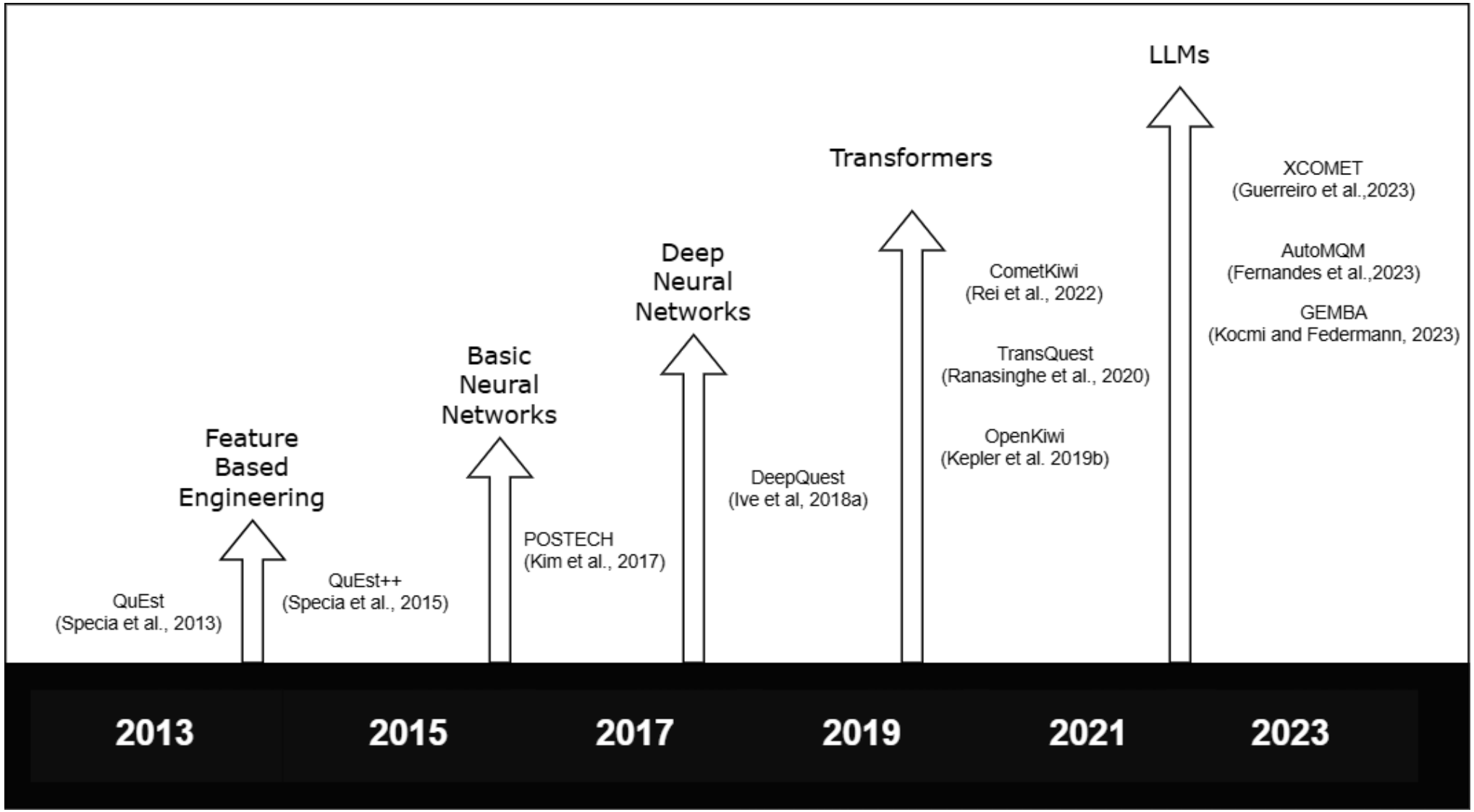
People-Centred AI
UNIVERSITY OF SURREY

CENTRE FOR TRANSLATION STUDIES
UNIVERSITY OF SURREY

Introduction

- *Evaluating* machine translation is a challenging cross-lingual task
  - *helps identify the reliability* of the translation and towards improvement of translation systems.
- Traditional Machine Translation evaluation methods are Reference-based which are resource-intensive.
- Quality estimation (QE) aims to score the quality of a translated text without a reference translation
  - Reduces cost and effort.
  - At different granularity levels – Segment-level, Word-level, Error Span Annotation, MQM

Quality Estimation *vis-à-vis* Language Modelling



Segment-level Quality Estimation

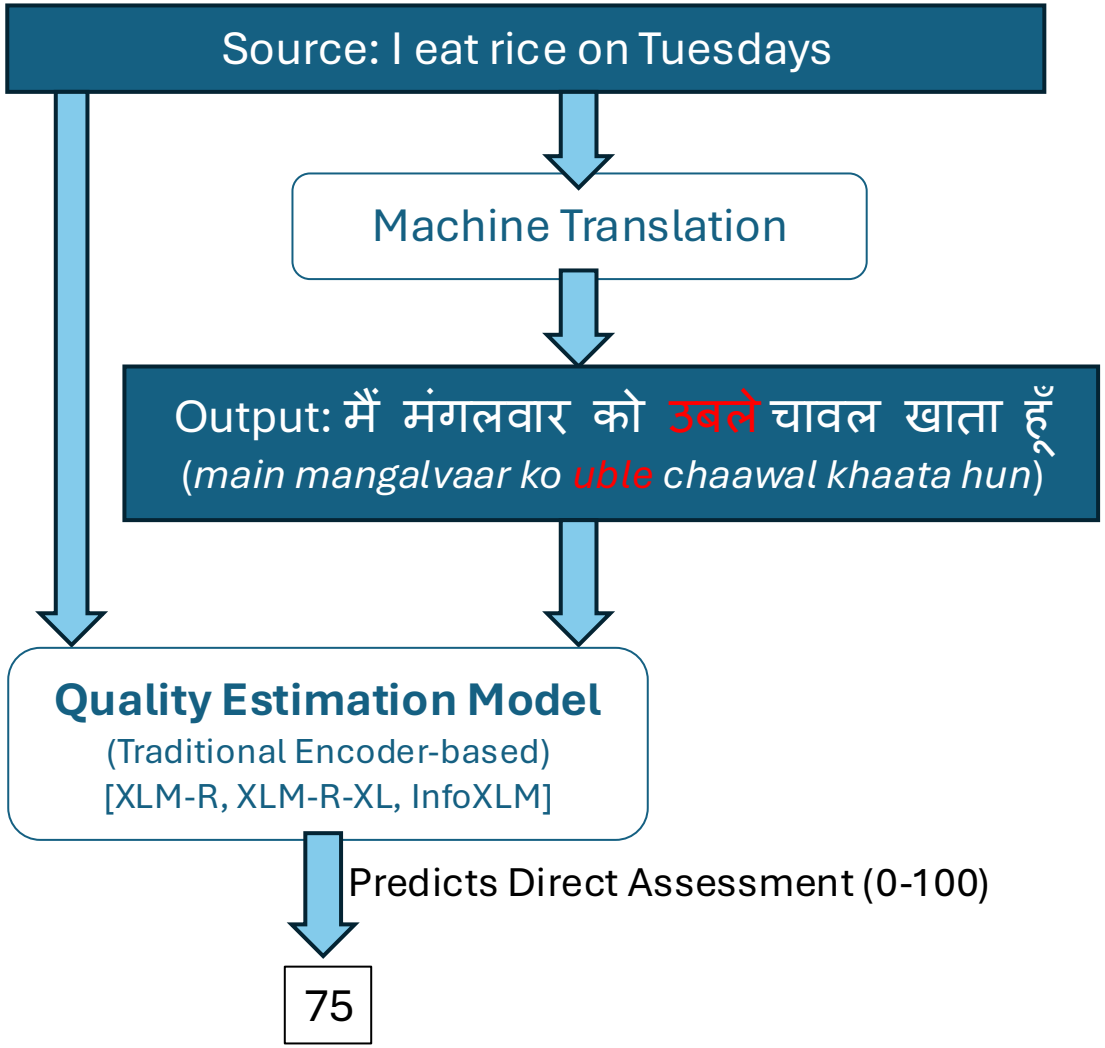
To provide a reliable, automatic measure of translation quality, crucial for system development and user-facing applications – without using a reference.

- Metrics like BLEU, chrF, MetricX **need a reference**.
- MT is subjective – multiple references – free order.

Quality Estimation (QE) is task of assessing the quality of machine-translated text in the absence of a human reference translation.

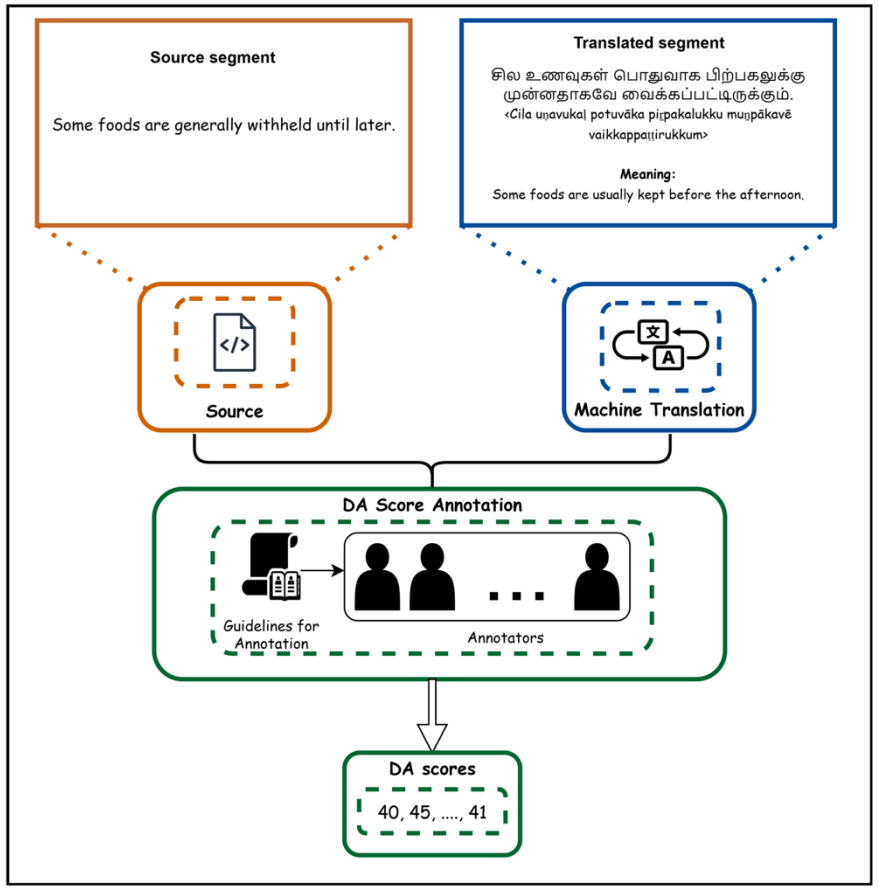
- **Segment-Level QE** focuses on assigning a quality score to a translated sentence, typically a Direct Assessment (DA) score from 0-100.

- **Word-Level QE** focuses on tagging each token in source and MT output with a OK/BAD tag, given the translation errors.



Direct Assessment (DA) score

Direct Assessment scores machine-translated content by having human assessors rate its quality (0 to 100)



Overall Score	Translation conveys source meaning?	How much translation conveys to source?
1 - 10	**Completely inaccurate.**	The MT output is unintelligible. Studying the meaning of the sentence is hopeless; even allowing for context, one feels that guessing would be too unreliable. ● The translation is incomprehensible, and machine translated output contains a mix of languages/dialects [which are not in the target language] (Adequacy) ● None of the keywords are translated in the target language. (Adequacy) ● There are major grammatical errors and typos (Fluency)
11 - 30	**Inaccurate but contains some keywords.**	● Incomprehensible translation (Fluency) ● Translation contains some keywords but not all (Adequacy) ● There are numerous grammatical errors and typos.
31 -50	**Partially. Target reflects partially the source.**	The general idea of the MT output is intelligible only after considerable study. ● Translation is only partially understandable, and the overall meaning is not conveyed. (Adequacy and Fluency) ● Translation contains some keywords (Adequacy) ● There are many grammatical errors and typos (Fluency)
51 - 70	**Yes. Target reflects the overall meaning.**	The MT output is generally clear and intelligible. Despite some inaccuracies or infelicities of the sentence, one can understand (almost) immediately what it means. ● Translation is understandable and reflects the source meaning. (Fluency) ● Translation contains most keywords (Adequacy) ● Only minor grammar errors (Fluency)
71 - 90	**Yes. Target reflects the source meaning without errors.**	The MT output is perfectly clear and intelligible. It is grammatical and reads like ordinary text. ● Translation is very closed to the source meaning (Fluency) ● Translation contains all keywords (Adequacy) ● No errors but there are better word choices in the target language. (Adequacy)
91 - 100	**Yes. Target reflects source meaning without errors.**	● Perfect translation (Adequacy and Fluency) ● Accurately reflects the meaning of the source (Fluency) ● No errors

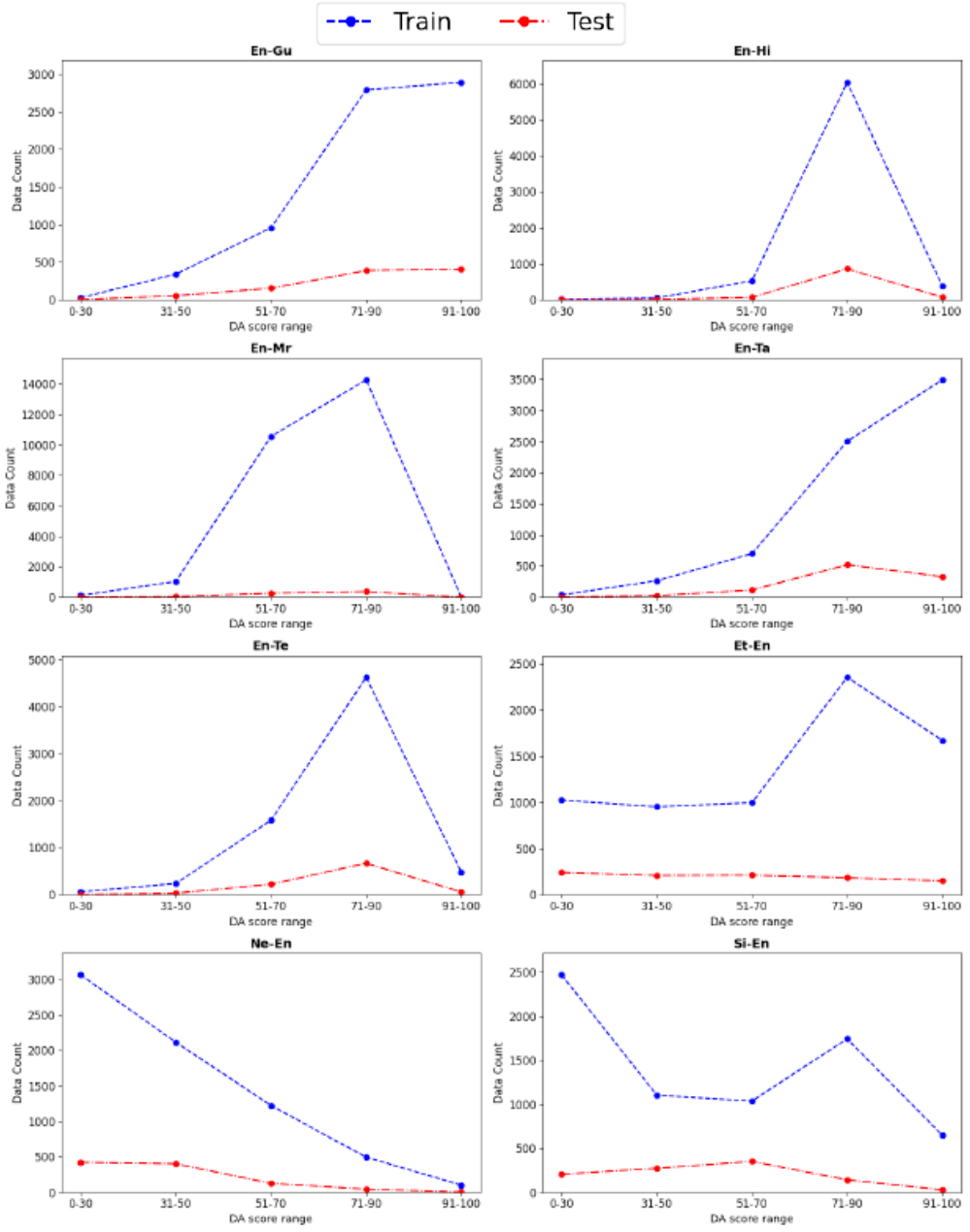
Dataset

- Focus on low-resource language pairs for QE.

- We utilize the DA score dataset from WMT QE shared task for our experiments.
  - En-Gu : English to ગુજરાતી (Gujarati)
  - En-Hi : English to **हिन्दी** (Hindi)
  - En-Mr : English to **मराठी** (Marathi)
  - En-Ta : English to **தமிழ்** (Tamil)
  - En-Te : English to **తెలుగు** (Telugu)

  - Et-En : **eesti keel** (Estonian) to English
  - Ne-En : **नेपाली** (Nepali) to English
  - Si-En : සිංහල (Sinhala) to English

Data Split	En-Gu	En-Hi	En-Mr	En-Ta	En-Te	Et-En	Ne-En	Si-En
Training	7000	7000	26,000	7000	7000	7000	7000	7000
Testing	1000	1000	699	1000	1000	1000	1000	1000



Meta-Evaluation for QE

- Spearman's Correlation **[Primary]**
- Pearson's Correlation
- Kendall's Tau
- Soft Pairwise Accuracy (SPA)

Applied **Williams test** to compare correlation significance between models and toidentify whether top models outperform others **statistically.**

Encoder-based QE

- **TransQuest** – Framework to fine-tune Transformers-based pre-trained language models for QE (Ranasinghe et al., 2020)
  - MonoTransQuest
  - SiameseTransQuest
- **COMETKiwi** – Continually pre-trained and fine-tuned models for QE (Rei et al., 2020; Unbabel)
  - COMETKiwi-22 [550M parameters]
  - COMETKiwi-23 [3.5B parameters]
- **xCOMET** – Multi-stage training to perform sentence-level and error-span detection

  [XL variant - 3.5B params].

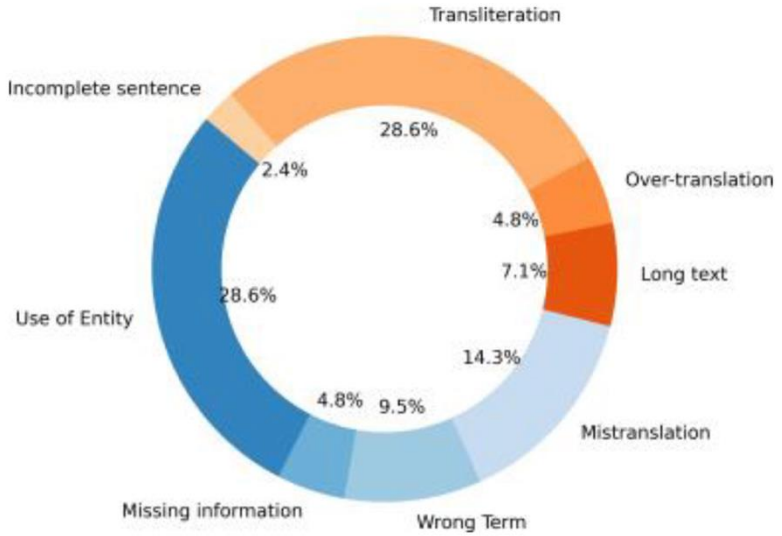
Encoder-based QE: *High- vs. Low-resource Languages*

Method	En-Gu	En-Hi	En-Mr	En-Ta	En-Te	Et-En	Ne-En	Si-En
MonoTQ-InfoXLM-Large	0.630	0.478	0.606	0.603	**0.358**	0.760	0.718	0.579
MonoTQ-XLM-V	0.552	0.446	0.556	0.607	0.358	0.766	0.705	0.595
MonoTQ-XLM-R-Large	0.599	0.491	0.457	**0.627**	0.339	0.737	0.700	0.575
CometKiwi-22 (InfoXLM)	0.574	0.408	0.627	0.567	0.231	0.827	0.755	0.648
CometKiwi-23 (XLM-R-XL)	**0.637**	**0.546**	**0.635**	0.616	0.338	**0.860**	**0.789**	**0.703**
xCOMET-XL	0.490	0.346	0.448	0.503	0.253	0.810	0.646	0.576

xCOMET-XL, and COMETKiwi use XLM-R-XL as the backbone while TransQuest fine-tunes on the chosen encoder model.

Performant for high-resource language pairs, however, target-side **low-resource languages suffer** due to:

- Morphological complexity (En-Ta, En-Te, En-Mr)
- *Meaningless* Tokenization
- Likely also, due to distribution of pre-training data for low-resource languages



Decoder / LLM-based QE

GEMBA (Kochmi et. al., 2023)

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:

Experimental Settings

- **Zero-Shot :** Model generates outputs using pre-trained knowledge and it's limited generalization ability.

- **Standard Instruction-Tuning :** Adapting a model using supervised fine-tuning that includes explicit instructions for QE task.

- **ALOPE :** Enhances fine-tuning by leveraging regression headers across multiple Transformer layers.

Models

- All experiments conducted with open-source LLMs with 8B parameters or less:
  - Llama-2-7B
  - Llama 3.1-8B
  - Llama 3.2-3B
  - Aya-expanse-8B

Challenges for LLM-based QE

- LLMs are **optimized for next-token prediction**, excelling at generative tasks but struggle with regression-based goals.
  - QE requires *numerical* and *cross-lingual reasoning*.
  - Limited ability to capture fine-grained relationships between language features and numerical scores.

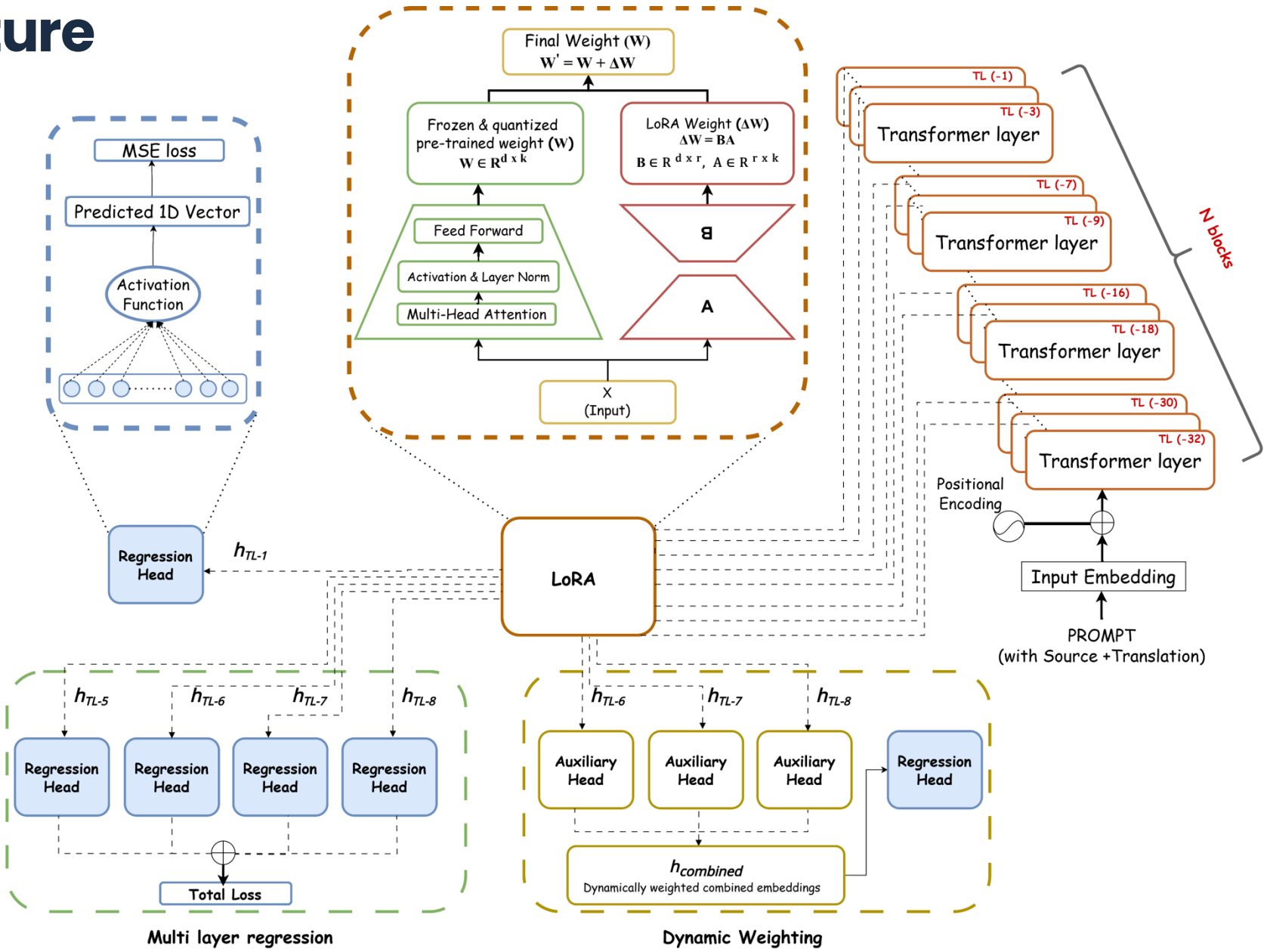
- Although LLMs refine context through multiple Transformer layers, standard practice usually predicts with the final layer for output.

ALOPE [Adaptive Layer Optimization for Translation Quality Estimation]

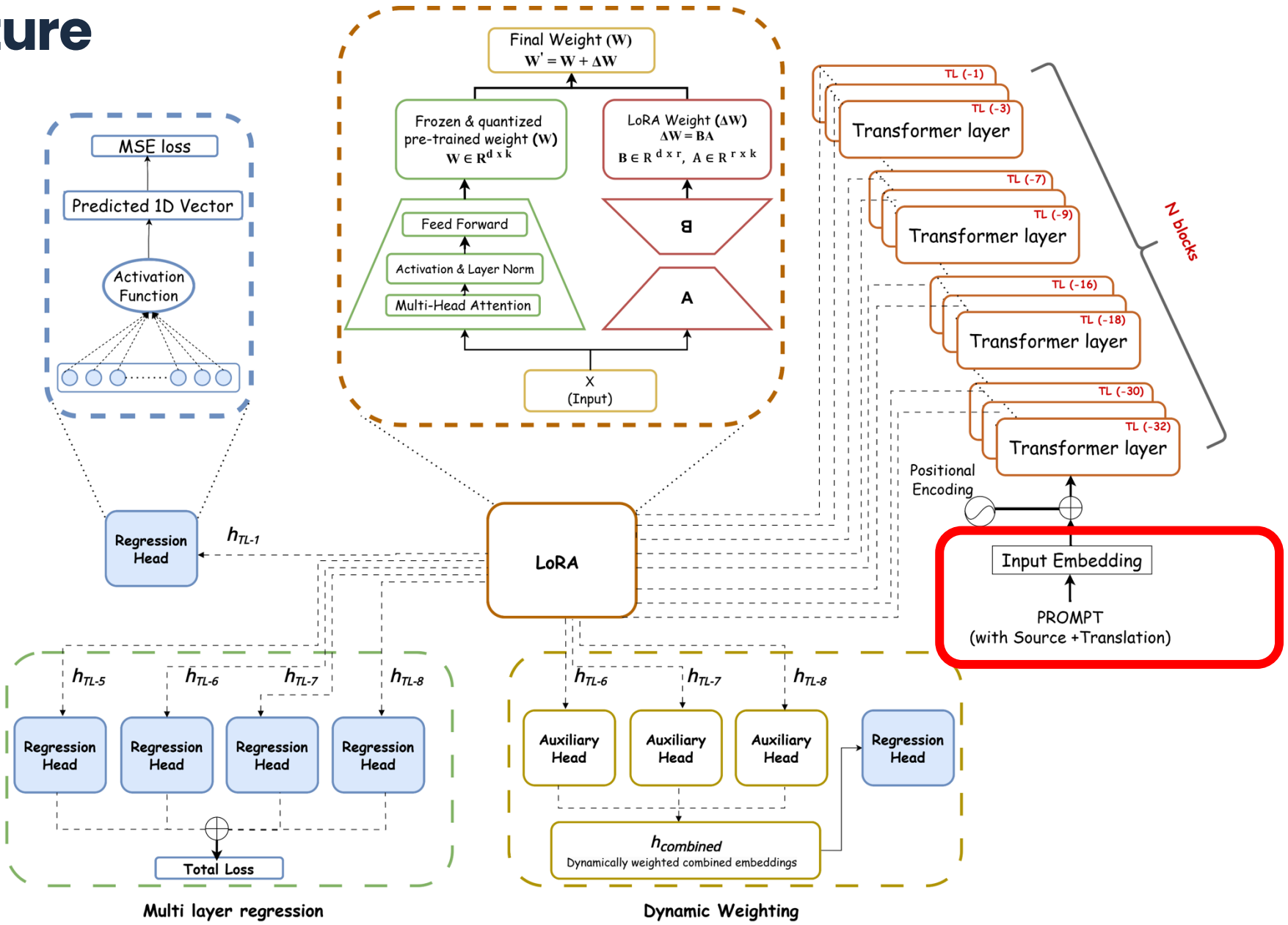
- ALOPE enables LLMs to
  - perform regression by integrating a novel head
  - uses a low-rank adapter with Transformer architecture.

- Helps analyze the impact of adaptive regression heads
  - to determine which Transformer's layers contribute most to LLM-based QE

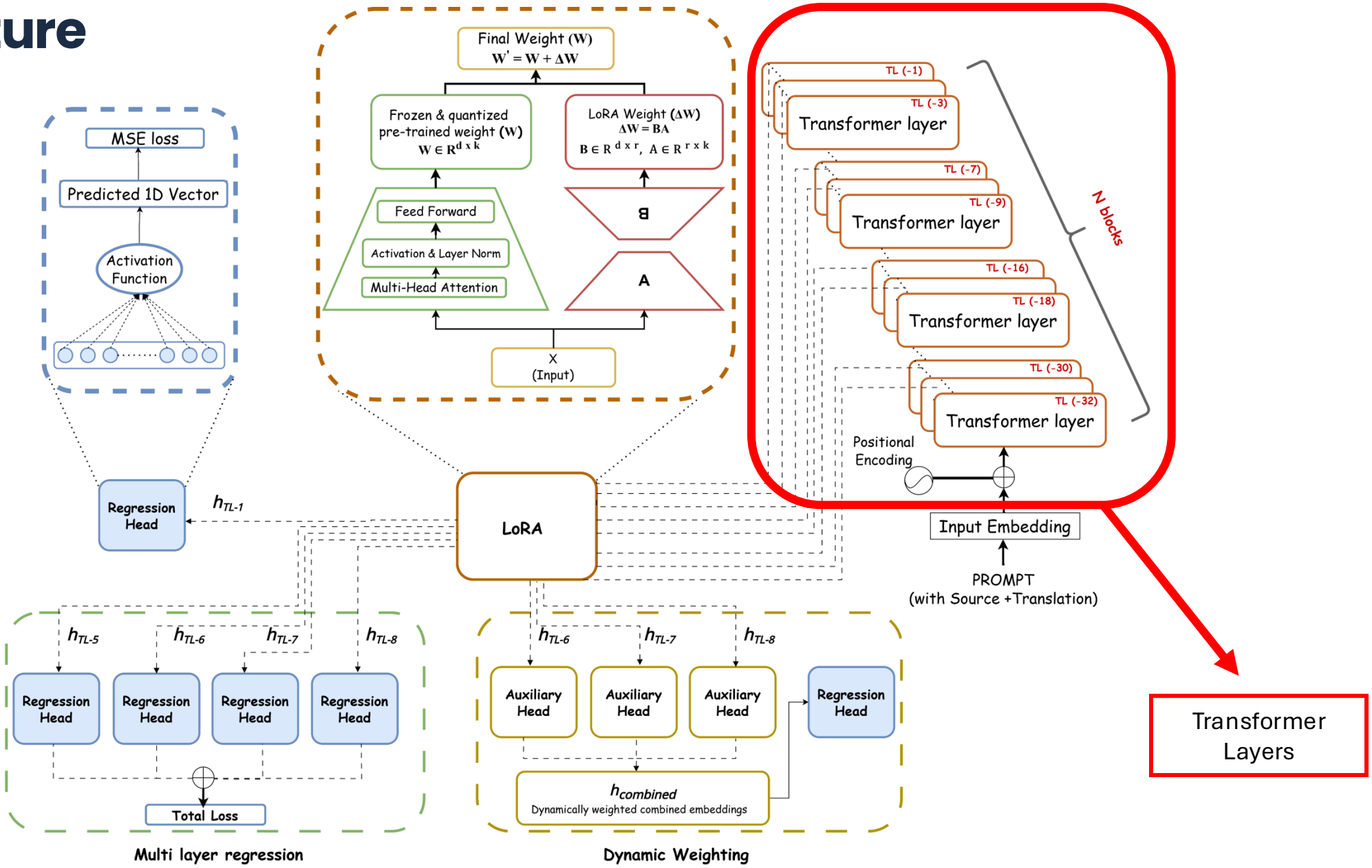
Architecture



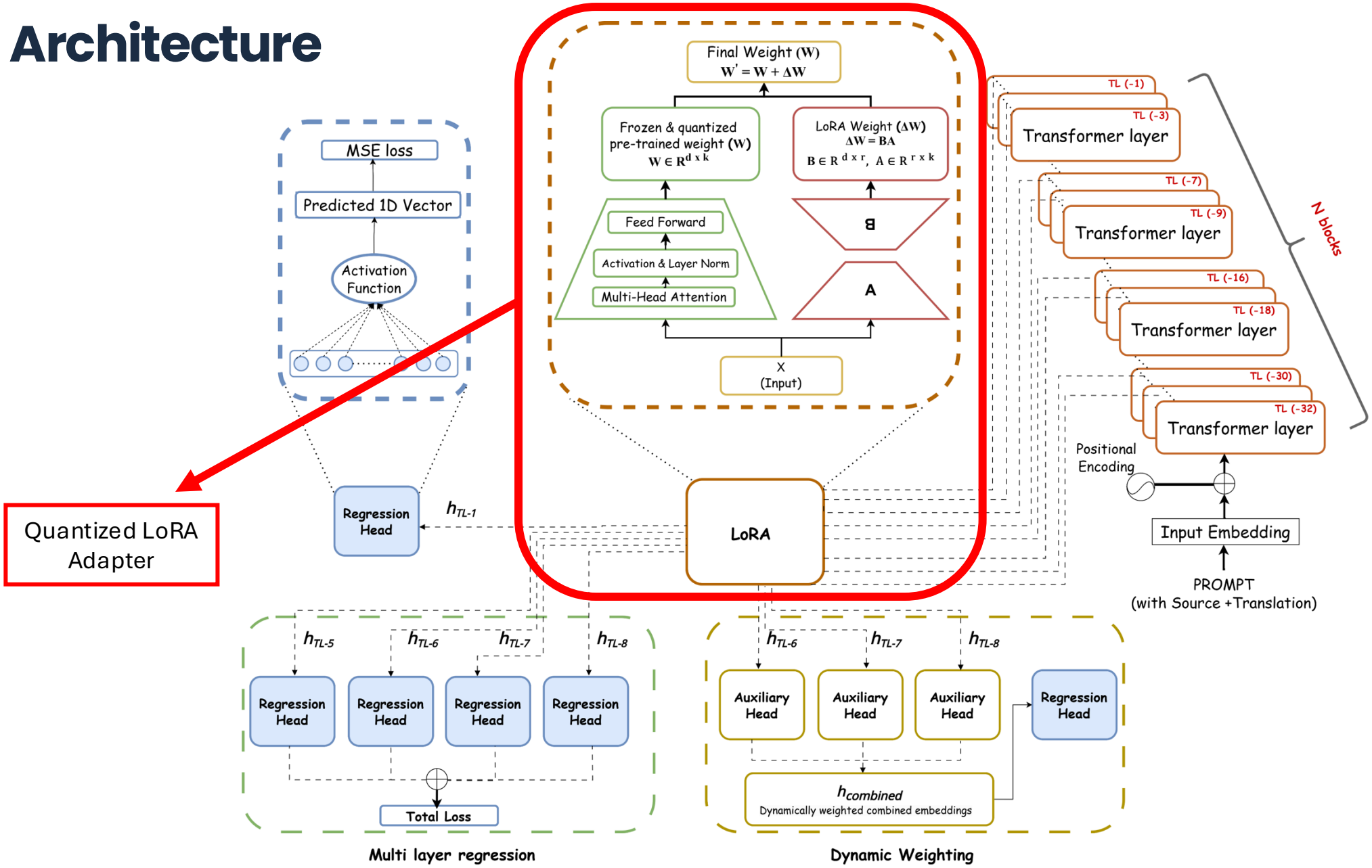
Architecture



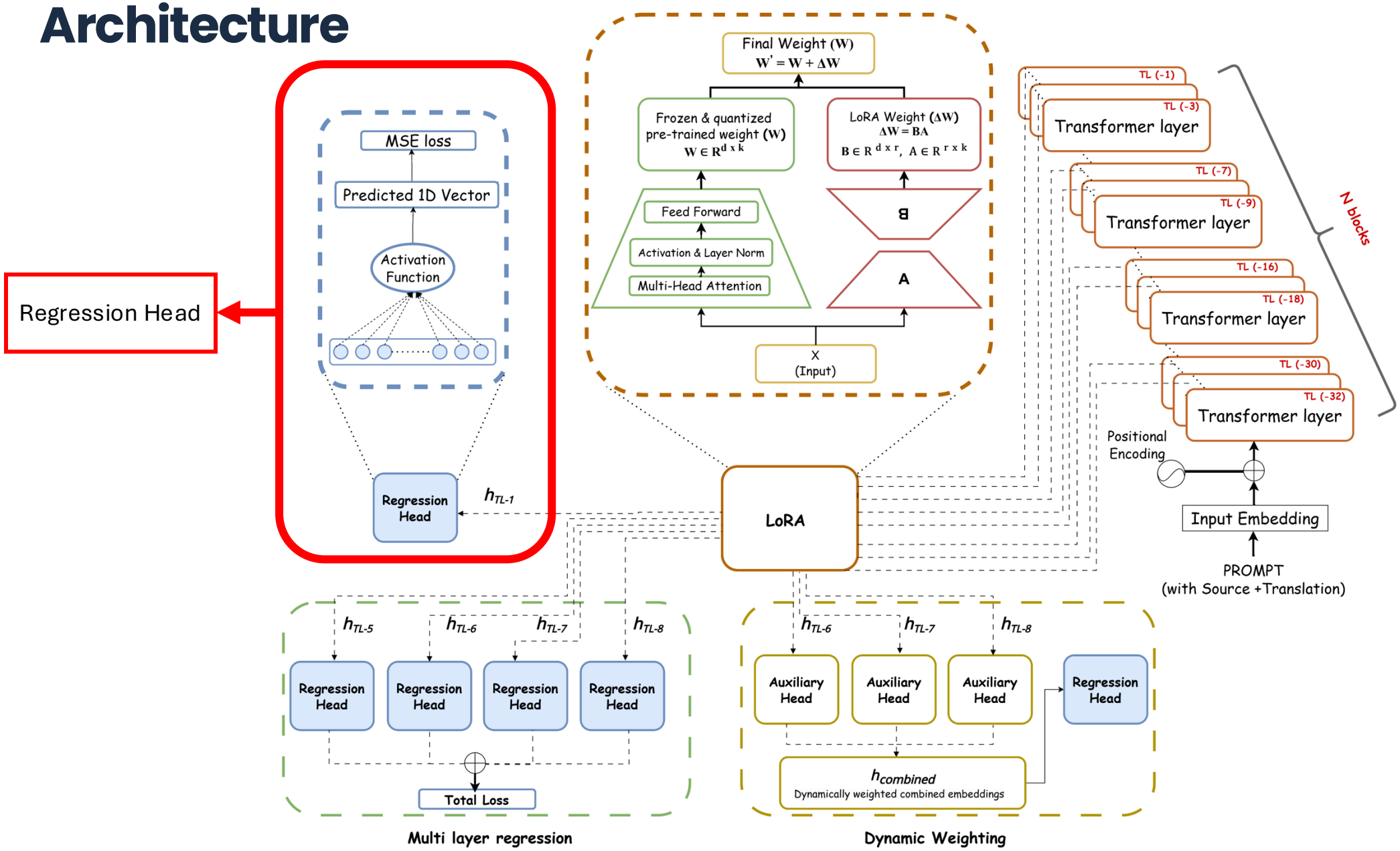
Architecture



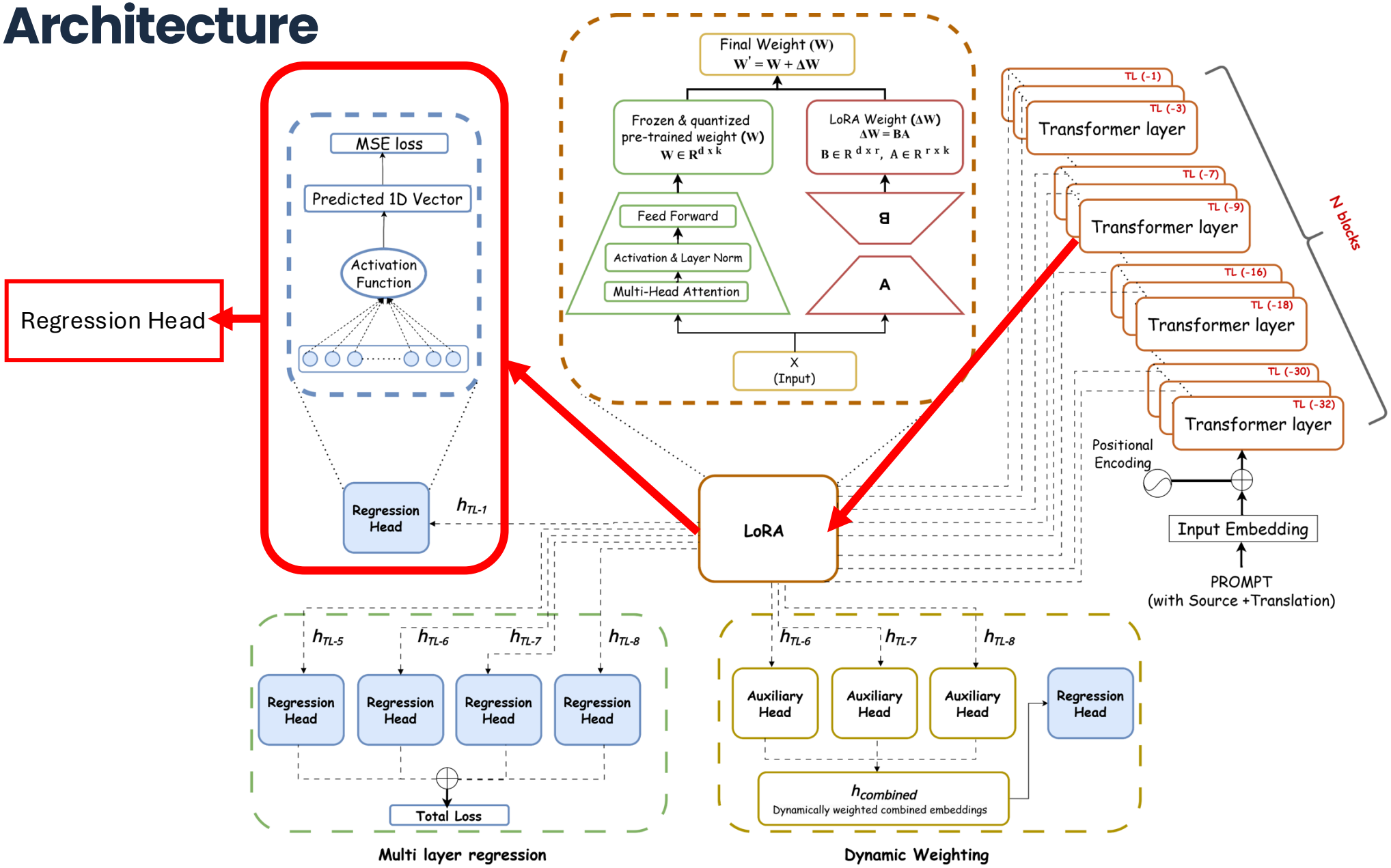
Architecture



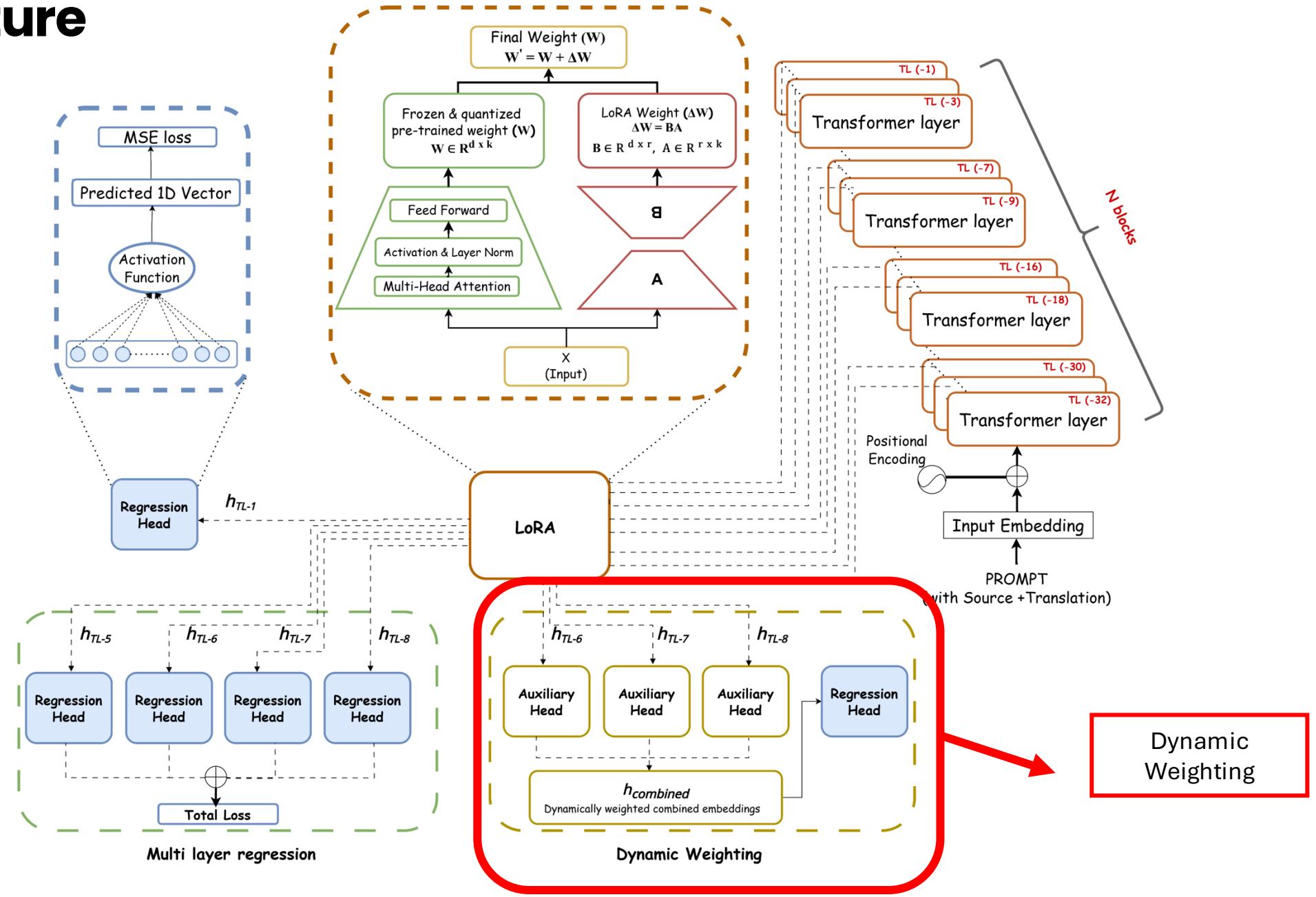
Architecture



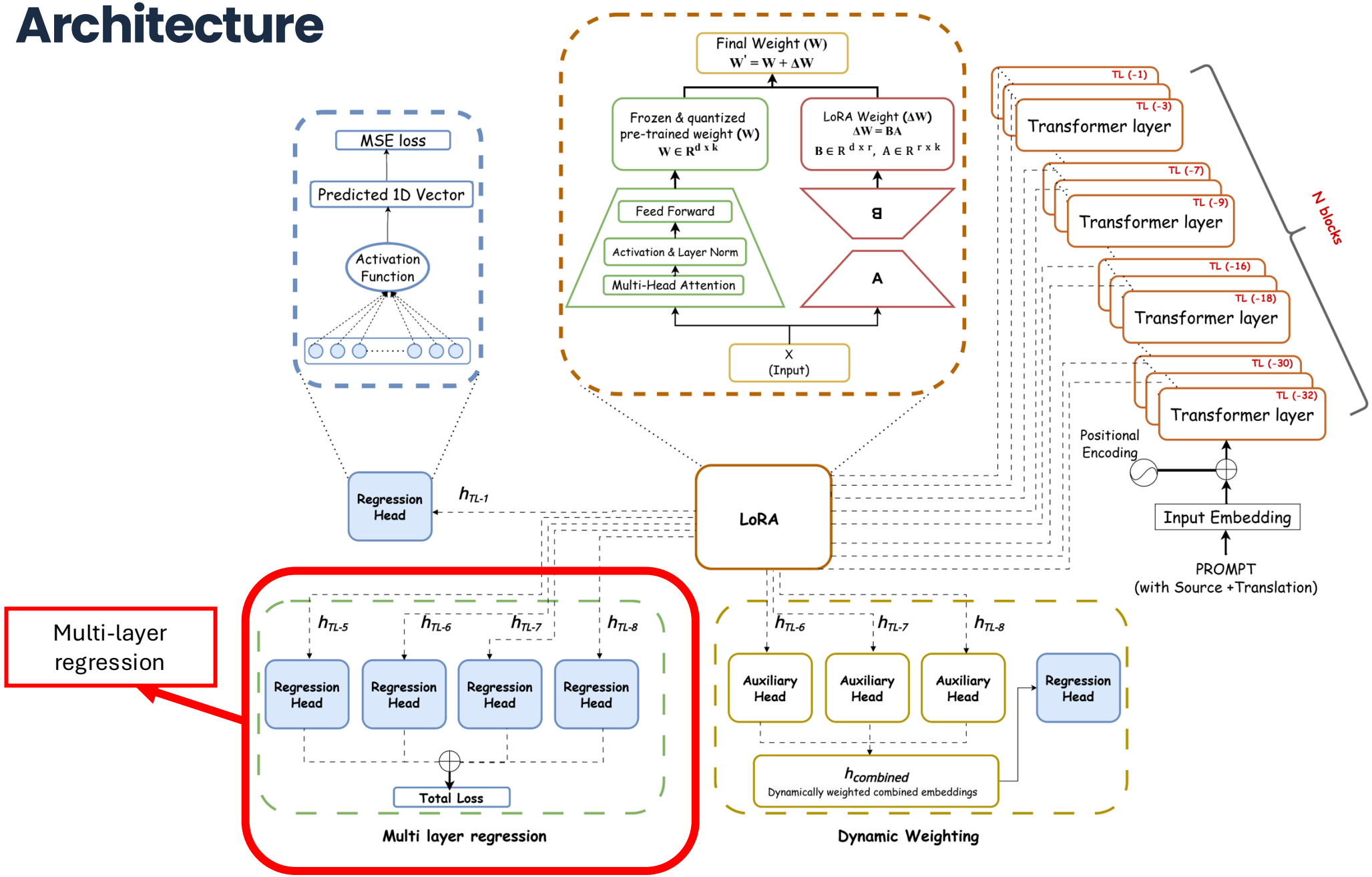
Architecture



Architecture



Architecture



ALOPE

Final Transformer Layer

Intermediate Transformer Layer

Intermediate Transformer Layer

Intermediate Transformer Layer

Lower-level Transformer Layer

Lower-level Transformer Layer

	Model	En-Gu	En-Hi	En-Mr	En-Ta	En-Te	Et-En	Ne-En	Si-En
TL (-1)	**llama-2-7b**	0.563	0.414	0.609	0.525	0.356	0.742	0.596	0.565
	llama 3.1-8B	0.594	0.469	0.620	0.567	0.363	0.734	0.647	0.547
	llama 3.2-3B	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)	**llama-2-7b**	0.567	0.336	0.542	0.484	0.317	0.739	0.606	**0.573**
	llama 3.1-8B	0.590	0.477	0.625	0.528	**0.388**	0.744	0.638	0.544
	llama 3.2-3B	**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)	**llama-2-7b**	0.360	0.301	0.361	0.254	0.293	0.405	0.164	0.049
	llama 3.1-8B	0.514	0.412	0.609	0.438	0.304	0.148	0.554	0.493
	llama 3.2-3B	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)	**llama-2-7b**	0.540	0.381	0.585	0.482	0.308	0.751	0.580	0.569
	llama 3.1-8B	0.558	0.453	0.602	0.523	0.350	0.737	0.652	0.513
	llama 3.2-3B	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
TL (-20)	**llama-2-7b**	0.470	0.405	0.544	0.460	0.338	0.684	0.508	0.534
	llama 3.1-8B	0.484	0.394	0.553	0.321	0.172	0.649	0.524	0.494
	llama 3.2-3B	0.430	0.408	0.579	0.303	0.286	0.601	0.488	0.464
	aya-expanse-8b	0.437	0.300	0.488	0.263	0.287	0.483	0.438	0.395
TL (-24)	**llama-2-7b**	0.500	0.421	0.538	0.379	0.239	0.630	0.507	0.472
	llama 3.1-8B	0.421	0.378	0.552	0.330	0.290	0.515	0.530	0.464
	llama 3.2-3B	0.443	0.376	0.507	0.367	0.299	0.559	0.528	0.487
	aya-expanse-8b	0.375	0.319	0.440	0.337	0.220	0.393	0.407	0.345

ALOPE

	Model	En-Gu	En-Hi	En-Mr	En-Ta	En-Te	Et-En	Ne-En	Si-En
TL (-1)	llama-2-7b	0.563	0.414	0.609	0.525	0.356	0.742	0.596	0.565
	llama 3.1-8B	0.594	0.469	0.620	0.567	0.363	0.734	0.647	0.547
	llama 3.2-3B	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)	llama-2-7b	0.567	0.336	0.542	0.484	0.317	0.739	0.606	**0.573**
	llama 3.1-8B	0.590	0.477	0.625	0.528	**0.388**	0.744	0.638	0.544
	llama 3.2-3B	**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)	llama-2-7b	0.360	0.301	0.361	0.254	0.293	0.405	0.164	0.049
	llama 3.1-8B	0.514	0.412	0.609	0.438	0.304	0.148	0.554	0.493
	llama 3.2-3B	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)	llama-2-7b	0.540	0.381	0.585	0.482	0.308	0.751	0.580	0.569
	llama 3.1-8B	0.558	0.453	0.602	0.523	0.350	0.737	0.652	0.513
	llama 3.2-3B	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
TL (-20)	llama-2-7b	0.470	0.405	0.544	0.460	0.338	0.684	0.508	0.534
	llama 3.1-8B	0.484	0.394	0.553	0.321	0.172	0.649	0.524	0.494
	llama 3.2-3B	0.430	0.408	0.579	0.303	0.286	0.601	0.488	0.464
	aya-expanse-8b	0.437	0.300	0.488	0.263	0.287	0.483	0.438	0.395
TL (-24)	llama-2-7b	0.500	0.421	0.538	0.379	0.239	0.630	0.507	0.472
	llama 3.1-8B	0.421	0.378	0.552	0.330	0.290	0.515	0.530	0.464
	llama 3.2-3B	0.443	0.376	0.507	0.367	0.299	0.559	0.528	0.487
	aya-expanse-8b	0.375	0.319	0.440	0.337	0.220	0.393	0.407	0.345

ALOPE

	Model	En-Gu	En-Hi	En-Mr	En-Ta	En-Te	Et-En	Ne-En	Si-En	*Avg.*
TL (-1)	llama-2-7b	0.563	0.414	0.609	0.525	0.356	0.742	0.596	0.565	*0.546*
	llama 3.1-8B	0.594	0.469	0.620	0.567	0.363	0.734	0.647	0.547	*0.567*
	llama 3.2-3B	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)	llama-2-7b	0.567	0.336	0.542	0.484	0.317	0.739	0.606	**0.573**	*0.520*
	llama 3.1-8B	0.590	0.477	0.625	0.528	**0.388**	0.744	0.638	0.544	*0.567*
	llama 3.2-3B	**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)	llama-2-7b	0.360	0.301	0.361	0.254	0.293	0.405	0.164	0.049	*0.273*
	llama 3.1-8B	0.514	0.412	0.609	0.438	0.304	0.148	0.554	0.493	*0.434*
	llama 3.2-3B	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)	llama-2-7b	0.540	0.381	0.585	0.482	0.308	0.751	0.580	0.569	*0.524*
	llama 3.1-8B	0.558	0.453	0.602	0.523	0.350	0.737	0.652	0.513	*0.548*
	llama 3.2-3B	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)	llama-2-7b	0.470	0.405	0.544	0.460	0.338	0.684	0.508	0.534	*0.493*
	llama 3.1-8B	0.484	0.394	0.553	0.321	0.172	0.649	0.524	0.494	*0.449*
	llama 3.2-3B	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)	llama-2-7b	0.500	0.421	0.538	0.379	0.239	0.630	0.507	0.472	*0.461*
	llama 3.1-8B	0.421	0.378	0.552	0.330	0.290	0.515	0.530	0.464	*0.435*
	llama 3.2-3B	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	

ALOPE

	Model	En-Gu	En-Hi	En-Mr	En-Ta	En-Te	Et-En	Ne-En	Si-En	*Avg.*
TL (-1)	llama-2-7b	0.563	0.414	0.609	0.525	0.356	0.742	0.596	0.565	*0.546*
	llama 3.1-8B	0.594	0.469	0.620	0.567	0.363	0.734	0.647	0.547	*0.567*
	llama 3.2-3B	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)	llama-2-7b	0.567	0.336	0.542	0.484	0.317	0.739	0.606	**0.573**	*0.520*
	llama 3.1-8B	0.590	0.477	0.625	0.528	**0.388**	0.744	0.638	0.544	*0.567*
	llama 3.2-3B	**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)	llama-2-7b	0.360	0.301	0.361	0.254	0.293	0.405	0.164	0.049	*0.273*
	llama 3.1-8B	0.514	0.412	0.609	0.438	0.304	0.148	0.554	0.493	*0.434*
	llama 3.2-3B	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)	llama-2-7b	0.540	0.381	0.585	0.482	0.308	0.751	0.580	0.569	*0.524*
	llama 3.1-8B	0.558	0.453	0.602	0.523	0.350	0.737	0.652	0.513	*0.548*
	llama 3.2-3B	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)	llama-2-7b	0.470	0.405	0.544	0.460	0.338	0.684	0.508	0.534	*0.493*
	llama 3.1-8B	0.484	0.394	0.553	0.321	0.172	0.649	0.524	0.494	*0.449*
	llama 3.2-3B	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)	llama-2-7b	0.500	0.421	0.538	0.379	0.239	0.630	0.507	0.472	*0.461*
	llama 3.1-8B	0.421	0.378	0.552	0.330	0.290	0.515	0.530	0.464	*0.435*
	llama 3.2-3B	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	

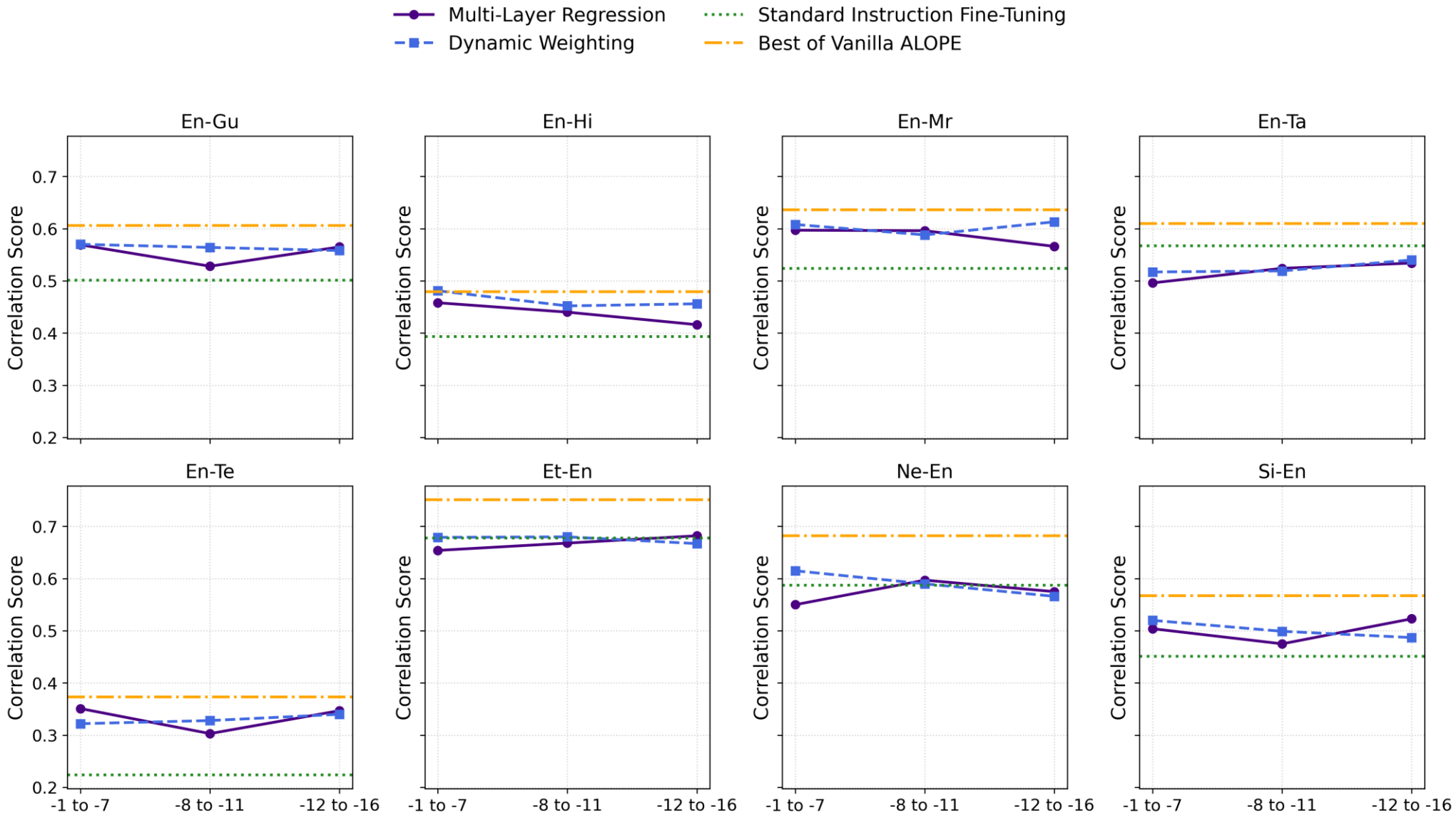
ALOPE

	Model	En-Gu	En-Hi	En-Mr	En-Ta	En-Te	Et-En	Ne-En	Si-En	*Avg.*
TL (-1)	llama-2-7b	0.563	0.414	0.609	0.525	0.356	0.742	0.596	0.565	*0.546*
	llama 3.1-8B	0.594	0.469	0.620	0.567	0.363	0.734	0.647	0.547	*0.567*
	llama 3.2-3B	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)	llama-2-7b	0.567	0.336	0.542	0.484	0.317	0.739	0.606	**0.573**	*0.520*
	llama 3.1-8B	0.590	0.477	0.625	0.528	**0.388**	0.744	0.638	0.544	*0.567*
	llama 3.2-3B	**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)	llama-2-7b	0.360	0.301	0.361	0.254	0.293	0.405	0.164	0.049	*0.273*
	llama 3.1-8B	0.514	0.412	0.609	0.438	0.304	0.148	0.554	0.493	*0.434*
	llama 3.2-3B	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)	llama-2-7b	0.540	0.381	0.585	0.482	0.308	0.751	0.580	0.569	*0.524*
	llama 3.1-8B	0.558	0.453	0.602	0.523	0.350	0.737	0.652	0.513	*0.548*
	llama 3.2-3B	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)	llama-2-7b	0.470	0.405	0.544	0.460	0.338	0.684	0.508	0.534	*0.493*
	llama 3.1-8B	0.484	0.394	0.553	0.321	0.172	0.649	0.524	0.494	*0.449*
	llama 3.2-3B	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)	llama-2-7b	0.500	0.421	0.538	0.379	0.239	0.630	0.507	0.472	*0.461*
	llama 3.1-8B	0.421	0.378	0.552	0.330	0.290	0.515	0.530	0.464	*0.435*
	llama 3.2-3B	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	

ALOPE

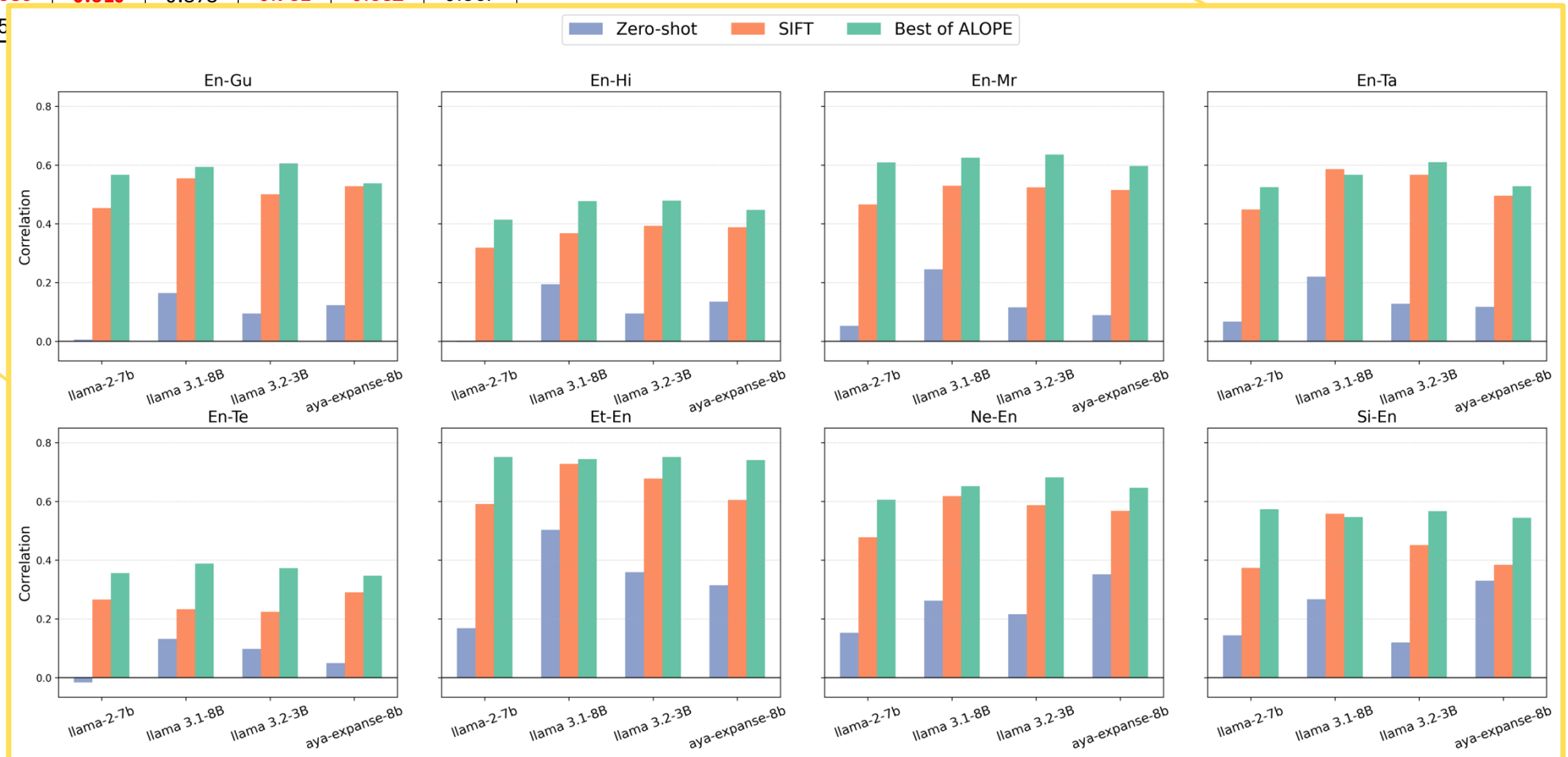
	Model	En-Gu	En-Hi	En-Mr	En-Ta	En-Te	Et-En	Ne-En	Si-En	*Avg.*
TL (-1)	llama-2-7b	0.563	0.414	0.609	0.525	0.356	0.742	0.596	0.565	*0.546*
	llama 3.1-8B	0.594	0.469	0.620	0.567	0.363	0.734	0.647	0.547	*0.567*
	llama 3.2-3B	0.604	0.477	**0.636**	0.580	0.348	0.735	0.674	0.543	*0.575*
	aya-expanse-8b	0.068	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)	llama-2-7b	0.567	0.336	0.542	0.484	0.317	0.739	0.606	**0.573**	*0.520*
	llama 3.1-8B	0.590	0.477	0.625	0.528	**0.388**	0.744	0.638	0.544	*0.567*
	llama 3.2-3B	**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)	llama-2-7b	0.360	0.301	0.361	0.254	0.293	0.405	0.164	0.049	*0.273*
	llama 3.1-8B	0.514	0.412	0.609	0.438	0.304	0.148	0.554	0.493	*0.434*
	llama 3.2-3B	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)	llama-2-7b	0.540	0.381	0.585	0.482	0.308	0.751	0.580	0.569	*0.524*
	llama 3.1-8B	0.558	0.453	0.602	0.523	0.350	0.737	0.652	0.513	*0.548*
	llama 3.2-3B	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)	llama-2-7b	0.470	0.405	0.544	0.460	0.338	0.684	0.508	0.534	*0.493*
	llama 3.1-8B	0.484	0.394	0.553	0.321	0.172	0.649	0.524	0.494	*0.449*
	llama 3.2-3B	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)	llama-2-7b	0.500	0.421	0.538	0.379	0.239	0.630	0.507	0.472	*0.461*
	llama 3.1-8B	0.421	0.378	0.552	0.330	0.290	0.515	0.530	0.464	*0.435*
	llama 3.2-3B	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	

Approaches with ALOPE vs SFIT



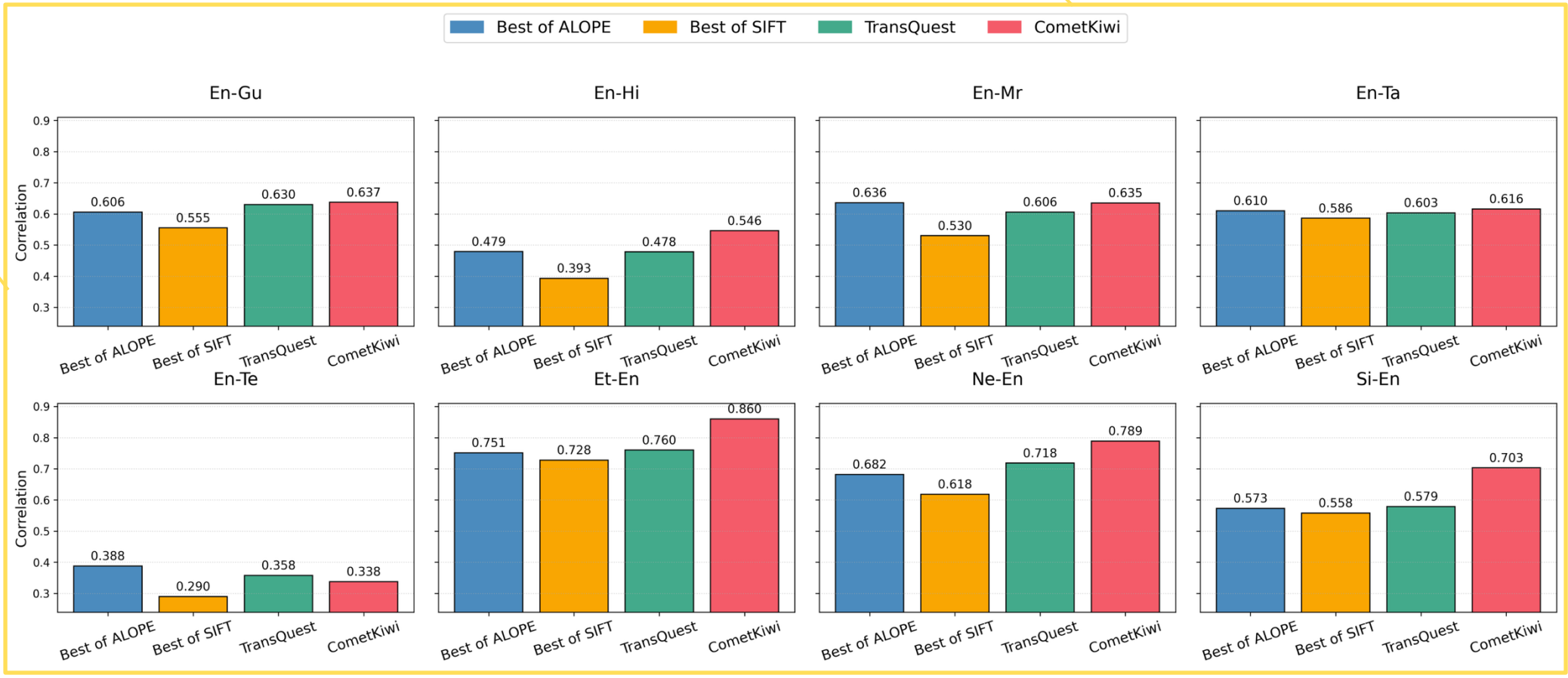
ALOPE *vs.* Zero-shot *vs.* SIFT

	Model	En-Gu	En-Hi	En-Mr	En-Ta	En-Te	Et-En	Ne-En	Si-En
Zero-shot	llama-2-7b	0.006	-0.002	0.053	0.067	-0.016	0.168	0.153	0.144
	llama 3.1-8B	**0.164**	**0.194**	**0.245**	**0.220**	**0.132**	**0.503**	0.262	0.267
	llama 3.2-3B	0.095	0.095	0.116	0.128	0.098	0.359	0.216	0.120
	aya-expanse-8B	0.123	0.135	0.089	0.117	0.049	0.315	**0.352**	**0.330**
SIFT	llama-2-7b	0.454	0.319	0.466	0.449	0.266	0.591	0.478	0.374
	llama 3.1-8B	**0.555**	0.368	**0.530**	**0.586**	0.233	**0.728**	**0.618**	**0.558**
	llama 3.2-3B	0.501	**0.393**	0.524	0.567	0.224	0.678	0.587	0.451
	aya-expanse-8B	0.528	0.388	0.515	0.496	**0.290**	0.605	0.568	0.384
Best of ALOPE	llama-2-7b	0.567	0.414	0.609	0.525	0.356	0.751	0.606	**0.573**
	llama 3.1-8B	0.594	0.477	0.625	0.567	**0.388**	0.744	0.652	0.547
	llama 3.2-3B	**0.606**	**0.479**	**0.636**	**0.610**	0.373	**0.751**	**0.682**	0.567
	aya-expanse-8B	0.538	0.447	0.5					

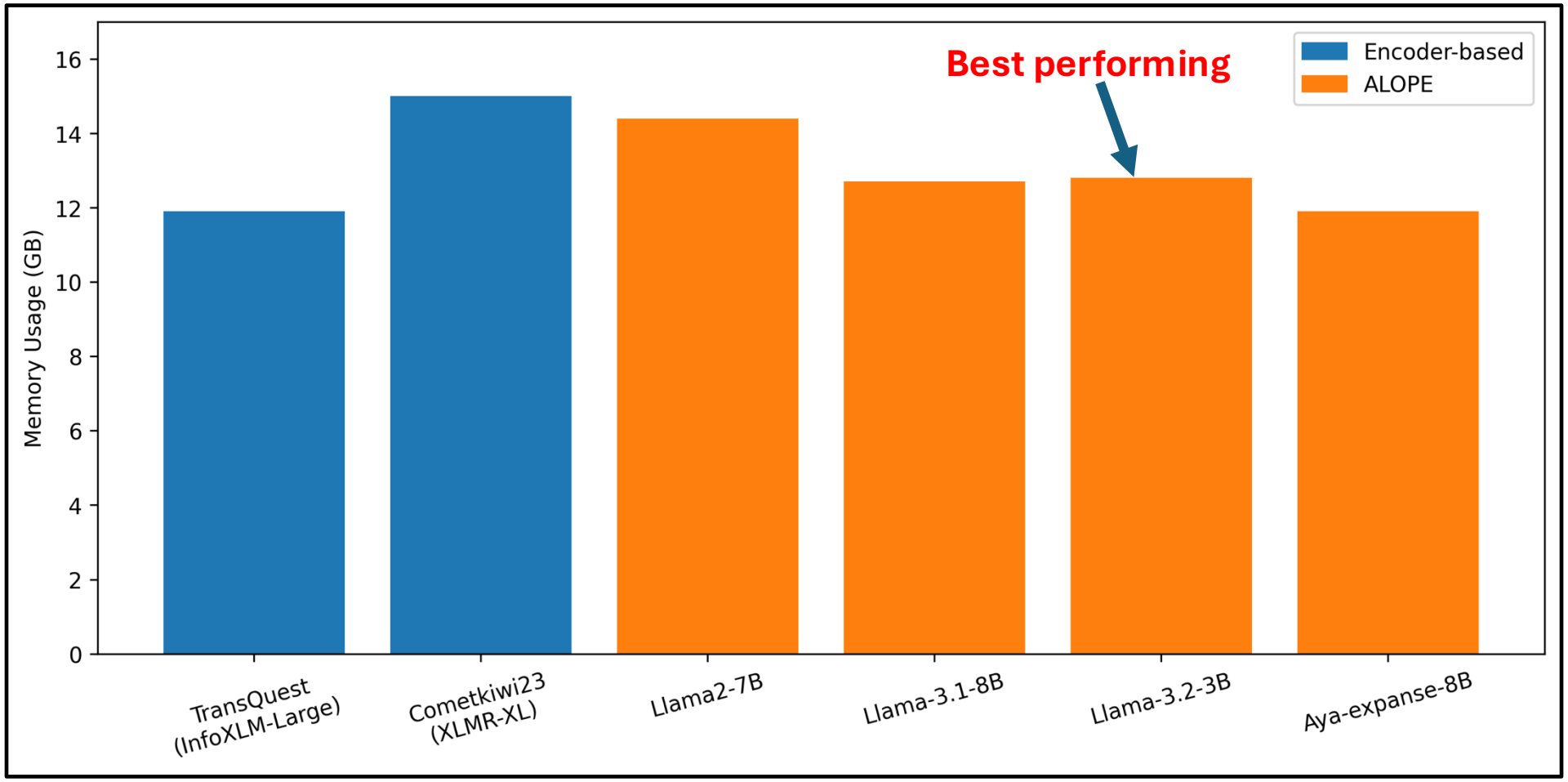


ALOPE *vs.* SIFT *vs.* SOTA Approaches

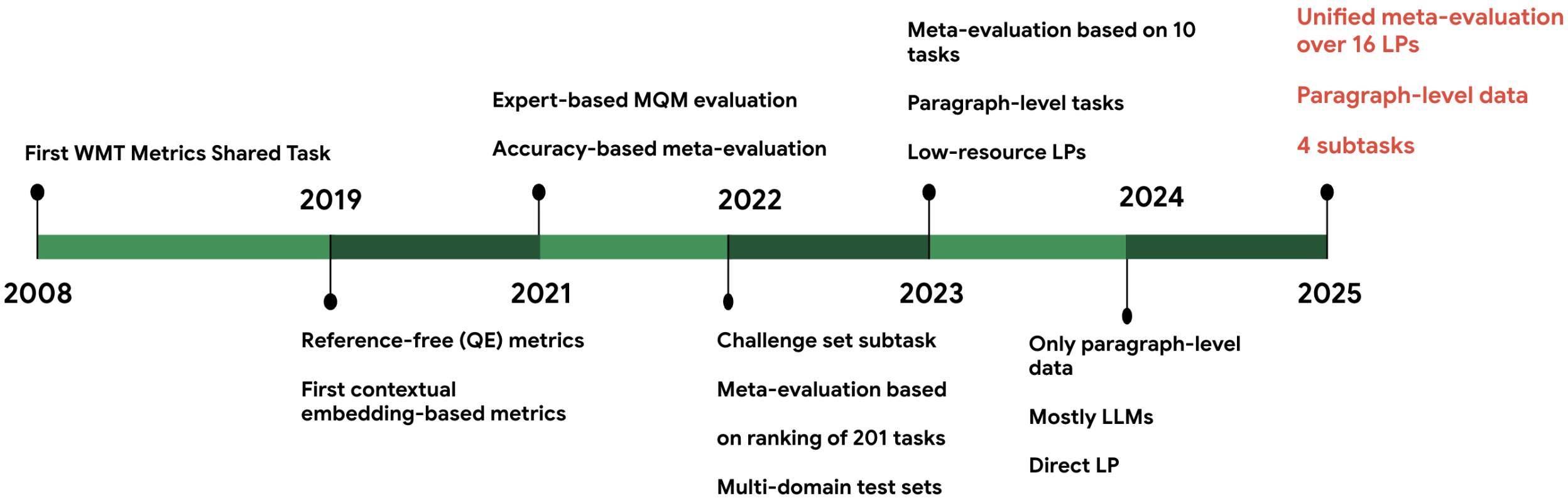
Approach	En-Gu	En-Hi	En-Mr	En-Ta	En-Te	Et-En	Ne-En	Si-En
Best ALOPE	0.606	0.479	**0.636**	0.610	**0.388**	0.751	0.682	0.573
Best from SIFT	0.555	0.393	0.530	0.586	0.290	0.728	0.618	0.558
TransQuest	0.630	0.478	0.606	0.603	0.358	0.760	0.718	0.579
CometKiwi	**0.637**	**0.546**	0.635	**0.616**	0.338	**0.860**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



Computational Efficiency



A Unified Metrics/QE/APE Shared Task



Subtask 3 aims to correct translation based on QE score

Task 1: Segment-level quality score prediction
Given source-hypothesis pair and optionally a reference segment, predict a quality score for each example in the evaluation set

Task 2: Span-level error detection
Given source-hypothesis pair and optionally a reference segment, predict the precise span of each translation error along with its severity

Task 3: Quality-informed segment-level error correction
Given source-hypothesis pair, automatically generated error spans, and optionally a reference segment, improve the translation

Task 4: Challenge sets
External "breakers" provide challenging test sets and perform challenge-focused evaluation of the participants in tasks 1 and 2

Source: Non parliamo italiano.

Reference: We don't speak Italian.

Hypothesis: I don't speak Spanish!

Post-edit: We don't speak Italian.

Task 3 aims to correct based on QE score

Improvements made by correction systems measured using COMET (Rei et. al., 2020) via

- **ΔCOMET** = COMET(source, corrected translation) - COMET(source, original translation)

We also measure **efficiency of corrections** using

- **Gain-to-Edit Ratio (GER)** = ΔCOMET / TER(original translation, corrected translation)

Additionally, we measure system performance using

- **ΔBLEURT** (BLEURT; Sellam et. al., 2020) and **ΔChrF++*** (ChrF++; Popović, 2017) to observe **semantic,** and **lexical similarity** with the provided reference, and
- **Batchwise Significance Testing** observing system performance over random samples

*For En-Ja and En-Zh, we use ΔChrF

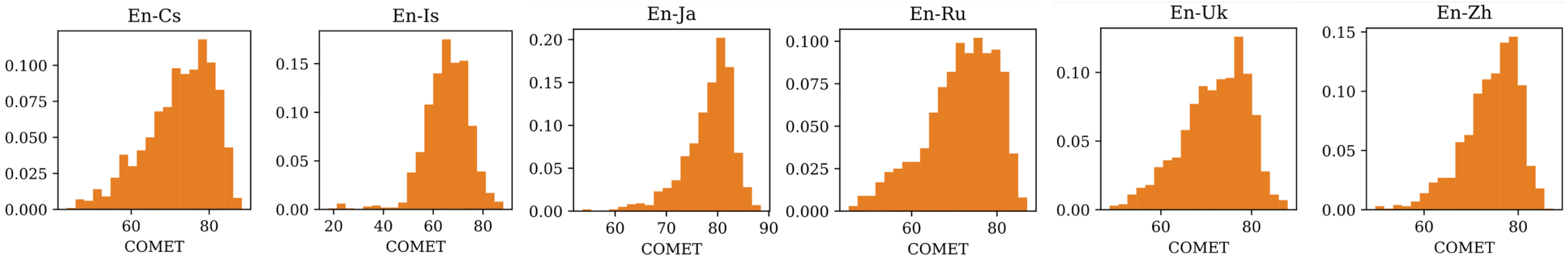
Task 3 Data and Baselines

We focus on **six language pairs** with **1,000 test instances *per* pair**.

- En-Cs, En-Is, En-Ja, En-Ru, En-Uk, En-Zh

Baselines

- **Translate from scratch** using Gemma3-IT-27B
- **Uses QE** output from XCOMET, **and post-edits** using Gemma3-IT-27B



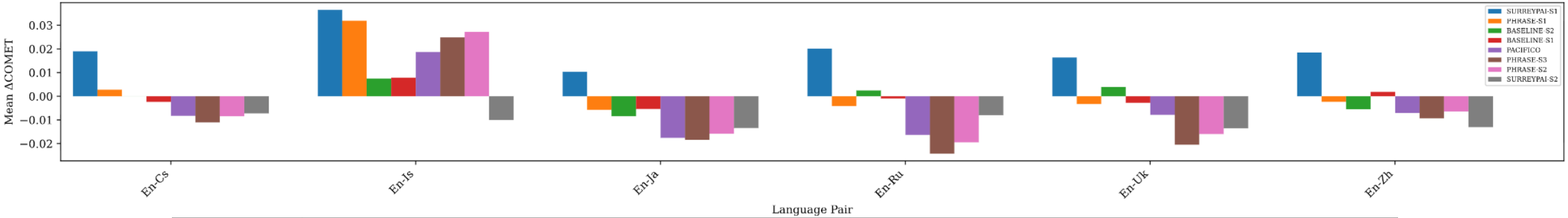
Task 3 Participants

Phrase (Hrabal et. al., 2025) leverage GPT-o3 and -o3 mini to correct outputs using *training-free approaches* based on *identifying fluency issues,* both, with and without reasoning steps (S1/S2), and *only via correcting errors* (S3).

SurreyPAI (Padmanabham et. al., 2025) used two training-free approaches, one which *leverages DA scores to identify an language-specific LLMs for re-translation*, and another which uses *fine-grained error-span information to post-edit* corrections.

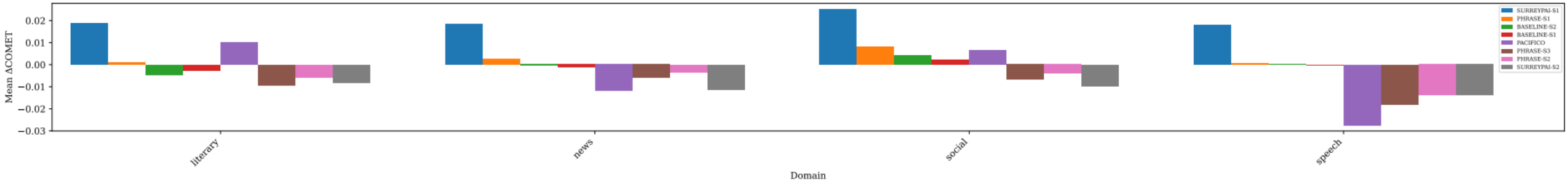
Pacifico (Sharma et. al., 2025) used explanations as an intermediate step in error detection and correction, using xTOWER to generate explanations, and then Gemini-1.5-pro for corrections.

Task 3: Primary Results



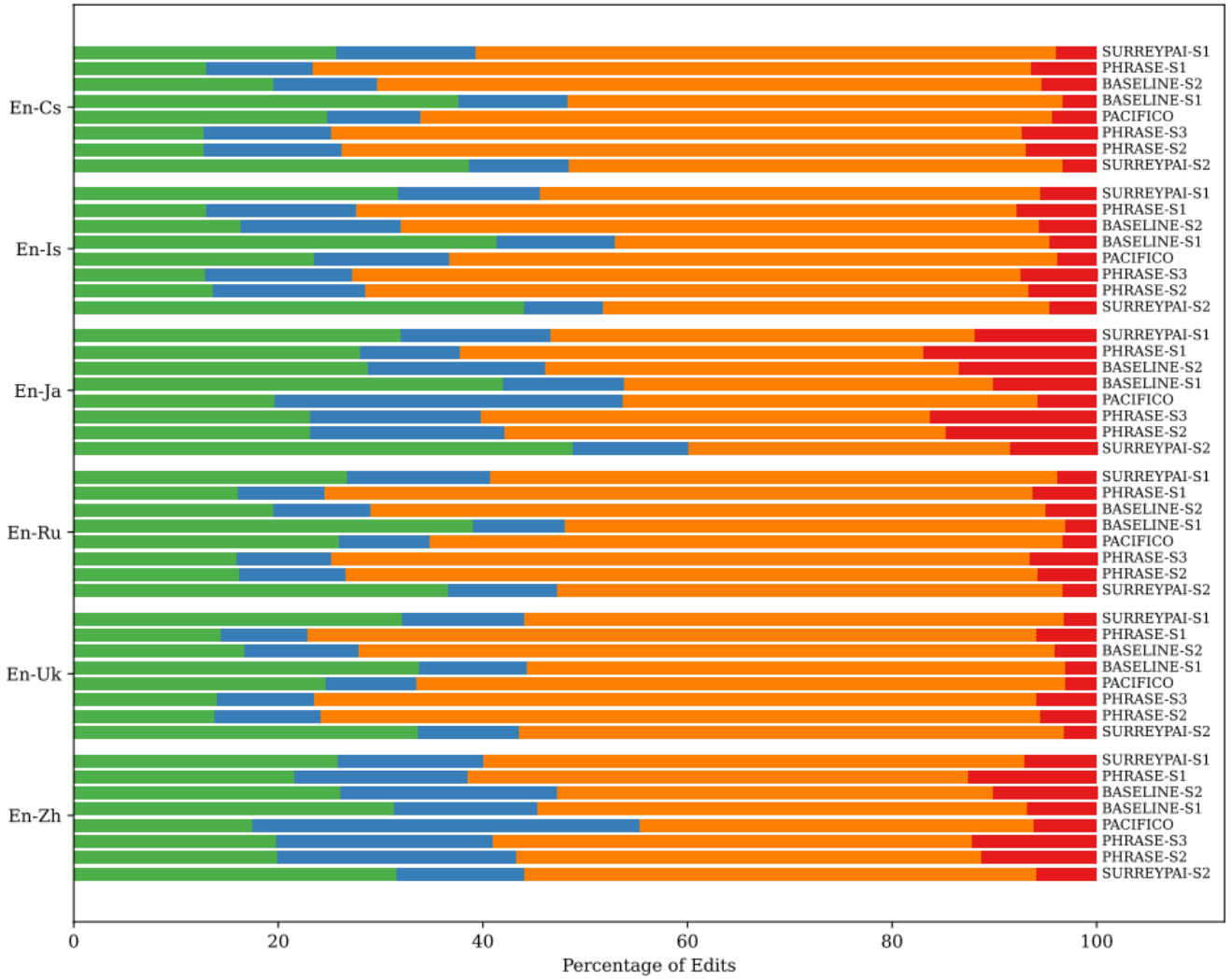
	ΔCOMET						GER					
	En-Cs	En-Is	En-Ja	En-Ru	En-Uk	En-Zh	En-Cs	En-Is	En-Ja*	En-Ru	En-Uk	En-Zh*
SurreyPAI-S1	**0.019**	**0.037**	**0.010**	**0.020**	**0.016**	**0.018**	**0.015**	0.027	**0.008**	**0.016**	**0.012**	**0.015**
Phrase-S1	0.003	0.032	-0.006	-0.004	-0.003	-0.002	0.006	0.058	-0.012	-0.007	-0.006	-0.005
Baseline-S1	0.000	0.007	-0.008	0.002	0.004	-0.005	0.000	0.026	-0.036	0.009	0.017	-0.023
Baseline-S2	-0.002	0.008	-0.005	-0.001	-0.003	0.002	-0.002	0.005	-0.004	-0.001	-0.002	0.002
Pacifico	-0.008	0.019	-0.018	-0.016	-0.008	-0.007	-0.032	0.054	-0.085	-0.061	-0.034	-0.036
Phrase-S3	-0.008	0.027	-0.016	-0.019	-0.016	-0.006	-0.030	**0.063**	-0.060	-0.056	-0.045	-0.023
Phrase-S2	-0.011	0.025	-0.018	-0.024	-0.020	-0.009	-0.027	0.050	-0.049	-0.050	-0.043	-0.026
SurreyPAI-S2	-0.007	-0.010	-0.013	-0.008	-0.014	-0.013	-0.005	-0.006	-0.008	-0.005	-0.009	-0.010

Task 3 Analysis: Domain, ΔCOMET, ΔBLEURT, and ΔChrF++



	ΔBLEURT						ΔChrF / ΔChrF++					
	En-Cs	En-Is	En-Ja	En-Ru	En-Uk	En-Zh	En-Cs	En-Is	En-Ja*	En-Ru	En-Uk	En-Zh*
SurreyPAI-S1	-0.002	0.053	0.003	0.009	0.002	-0.003	-3.460	0.000	0.000	0.000	0.000	0.000
Phrase-S1	-0.012	0.031	-0.033	-0.030	0.023	-0.045	2.256	2.856	-0.558	-11.636	13.758	-0.279
Baseline-S1	-0.027	0.025	-0.007	-0.006	-0.002	-0.007	0.059	0.264	-2.272	0.733	0.214	-1.086
Baseline-S2	-0.019	0.012	-0.009	0.011	0.018	0.007	8.902	2.199	4.358	0.954	7.267	-1.662
Pacifico	-0.035	-0.004	-0.030	-0.039	-0.024	-0.010	8.546	10.434	2.934	-42.936	6.791	5.989
Phrase-S3	-0.110	-0.075	-0.053	-0.210	-0.171	-0.090	6.550	7.117	-0.579	-14.537	8.471	0.331
Phrase-S2	-0.104	-0.076	-0.050	-0.206	-0.176	-0.083	4.810	5.120	-2.228	-16.787	3.223	4.760
SurreyPAI-S2	-0.015	0.002	-0.024	-0.007	0.002	-0.024	-1.219	-5.248	0.029	0.041	-22.708	0.000

Task 3 Analysis: Edit Operations



Substitution → Most frequent (>50%). Indicates that **lexical choice** is the main challenge for translation systems.

Deletion, **Insertion** → Next most common operations

Shift → Least frequent. Suggesting that models produce **syntactically coherent translations**.

Consistent **pattern** across all systems **Japanese & Chinese** show slightly **more insertions/deletions**

Task 3: Key Takeaways

Correcting LLM-based Translation is Challenging

This was the first task focused on correcting high-quality LLM translations.

APE systems struggled to improve them, similar to early struggles of enc-dec neural APE systems with NMT output.

Main Error: Word Choice, Not Fluency

Substitution was the dominant edit (>50%), as syntax was already plausible .

The key challenge is **lexical choice**.

Challenge Domain: Speech

Systems **struggled** most **with the speech domain**.

Correcting text from multimodal sources may require different methods.

Conclusion

- **Clear gap in performance** for high-resource language on the target side *vs*. low-resource languages.
- **ALOPE: Novel framework**
  - Regression heads + LoRA on LLMs
- **Performance**
  - Outperforms SIFT; matches / exceeds the performance of SOTA approaches QE
- **Findings**
  - Intermediate layers (TL-7, TL-11) => **most effective for low-resource cross-lingual QE**
  - Mid=> **layers stabilize earlier when English is target**
  - ALOPE w/ LLaMA-3.2 => **best overall despite small size**
- **Extension to the framework**
  - Dynamic weighting & multi-head regression improve baseline SFIT
- **Efficiency**
  - Competitive GPU memory , practical & scalable
- **Future vision**
  - Enhancing ALOPE for error reasoning and automatic post-editing
  - Look out for **ALOPE-RL**, and **ALOPE-APE** branches of this work! ☺

Thank You

Questions?

References

Lavie, Alon, Greg Hanneman, Sweta Agrawal, Diptesh Kanojia, Chi-Kiu Lo, Vilém Zouhar, Frederic Blain et al. "Findings of the WMT25 shared task on automated translation evaluation systems: Linguistic diversity is challenging and references still help." In *Proceedings of the Tenth Conference on Machine Translation*, pp. 436-483. 2025.

Sindhujan, Archchana, Shenbin Qian, Chan Chi Chun Matthew, Constantin Orasan, and Diptesh Kanojia. "ALOPE: Adaptive Layer Optimization for Translation Quality Estimation using Large Language Models." In *Second Conference on Language Modeling*.

Sindhujan, Archchana, Diptesh Kanojia, Constantin Orasan, and Shenbin Qian. "When LLMs Struggle: Reference-less Translation Evaluation for Low-resource Languages." In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pp. 437-459. 2025.

Sindhujan, Archchana, Diptesh Kanojia, and Constantin Orăsan. "Reference-Less Evaluation of Machine Translation: Navigating Through the Resource-Scarce Scenarios." *Information* 16, no. 10 (2025): 916.

Qian, Shenbin, Archchana Sindhujan, Minnie Kabra, Diptesh Kanojia, Constantin Orasan, Tharindu Ranasinghe, and Frédéric Blain. "What do Large Language Models Need for Machine Translation Evaluation?."

Sindhujan, Archchana, 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.