Domain-Specific Quality Estimation for Machine Translation in Low-Resource Scenarios

Namrata Patil Gurav, Akashdeep Ranu, Archchana Sindhujan, Diptesh Kanojia

Institute for People-Centred AI,

University of Surrey, United Kingdom

{np00996, ar02258, a.sindhujan, d.kanojia}@surrey.ac.uk

Abstract

Quality Estimation (QE) is essential for assessing machine translation quality in reference-less settings, particularly for domain-specific and low-resource language scenarios. In this paper, we investigate sentence-level QE for English→Indic machine translation across four domains (`Healthcare`, `Legal`, `Tourism`, and `General`) and five language pairs. We systematically compare zero-shot, few-shot, and guideline-anchored prompting across selected closed-weight and open-weight LLMs. Findings indicate that while closed-weight models achieve strong performance *via* prompting alone, prompt-only approaches remain fragile for open-weight models, especially in high-risk domains. To address this, we adopt `ALOPE`, a framework for LLM-based QE which uses Low-Rank Adaptation with regression heads attached to selected intermediate Transformer layers. We also extend `ALOPE` with the recently proposed Low-Rank Multiplicative Adaptation (LoRMA) for this work. Our results show that intermediate-layer adaptation consistently improves QE performance, with gains in semantically complex domains, indicating a way ahead for robust QE in practical scenarios. We release code and domain-specific QE datasets publicly for further research[1].

1 Introduction

Quality Estimation (QE) enables the output of Machine Translation (MT) systems to be evaluated at scale without requiring reference translations (Zerva et al., 2022). Unlike traditional MT evaluation metrics such as BLEU or METEOR, QE directly predicts a quality score for a source–translation pair, making it especially suitable for real-world deployment scenarios where reference translations are unavailable. QE can be performed at multiple granularities, and our work focuses on segment-level QE, predicting Direct Assessment (DA) scores on a continuous scale of $(0 \leq x \leq 100)$ for the given translation (Graham et al., 2013). The ground-truth DA score is obtained by averaging the scores from three or more human annotators.

While fine-grained annotation frameworks such as Multi-dimensional Quality Metrics (MQM) (Lommel et al., 2013) provide detailed insights into translation errors, they impose a substantial cognitive and temporal burden on annotators. In practice, reliable QE enables translation systems to prioritise human intervention for high-risk outputs, which is particularly important in sensitive or specialised domains (Zerva et al., 2024). Despite the widespread adoption of neural machine translation, translation quality remains uneven across languages and domains, particularly outside high-resource, `General`-domain settings (Specia et al., 2018; Zhao et al., 2024). This disparity is especially pronounced for English→Indic language pairs (Sindhujan et al., 2025a), where rich morphology, frequent code-mixing, script variation, and limited availability of high-quality parallel data continue to pose persistent challenges for both MT and evaluation (Zhao et al., 2024).

Although MT output for `General` content is often fluent, translations in domain-specific settings such as `Healthcare`, `Legal`, and `Tourism` remain fragile, as models are less exposed to specialised terminology and domain-specific constructions during training (Specia et al., 2018). Even minor errors involving negation, numerical values, units, or specialised terminology can result in substantial meaning changes, with potentially serious real-world consequences, particularly in high-risk domains such as `Healthcare` and `Legal` (Sindhujan et al., 2025b). These limitations underscore the need for robust domain-aware QE mechanisms that can reliably identify problematic translations in the absence of reference translations before deployment

[1]https://github.com/surrey-nlp/ALOPE/tree/main/Domain-based-QE-with-ALOPE

(Zerva et al., 2024).

Recent advances in Large Language Models (LLMs) have enabled QE through prompt-based scoring of source–translation pairs, offering an alternative in low-resource and domain-specific settings where supervised QE data are limited (Brown et al., 2020). However, prior work shows that prompt-only LLM-based QE remains inferior to state-of-the-art encoder-based QE models, particularly for sentence-level regression tasks (Zerva et al., 2022; Sindhujan et al., 2025b). A key limitation is that LLMs are optimised for next-token prediction rather than regression-oriented objectives such as predicting DA scores. Prompting or instruction tuning alone does not introduce regression-specific training signals, often resulting in unstable predictions (Zhao et al., 2021; Sindhujan et al., 2025c). Furthermore, most LLM-based QE approaches rely solely on representations from the final Transformer layer, despite growing evidence that intermediate layers may better encode cross-lingual and semantic alignment for low-resource languages (Kargaran et al., 2025).

Our work addresses these limitations by investigating domain-aware QE for English→Indic translation language pairs through a dual-track evaluation: (i) systematic comparison of prompt-only approaches across closed-weight and open-weight LLM families, and (ii) lightweight, parameter-efficient fine-tuning based approach where open-weight prompt-only methods prove insufficient.

Building on the `ALOPE` framework (Sindhujan et al., 2025c), which demonstrated that intermediate Transformer layers encode more stable QE-relevant signals than final-layer representations, we extend this approach to domain-specific, low-resource English→Indic settings (§ 3.3). `ALOPE` attach regression heads to informative intermediate layers and update only a minimal parameter subset using Low-Rank Adaptors (Hu et al., 2022), maintaining computational efficiency while improving the QE performance where deployment constraints preclude closed-weight API access.

We conduct a comprehensive evaluation across four domains (`Healthcare`, `Legal`, `Tourism`, and `General`) and five Indic languages (Hindi, Marathi, Tamil, Telugu, and Gujarati). Our experiments systematically compare prompt-only baselines (zero-shot, few-shot, and guideline-anchored prompting) and `ALOPE`, analysing performance differences between closed-weight models and open-weight models. Critically, our study establishes when lightweight `ALOPE`-based methods provide value versus when strong prompting alone suffices, offering practical deployment guidance for resource-constrained QE scenarios. The **main contributions** of this work are as follows:

- We provide a rigorous comparison of prompt-only QE strategies across closed-weight and open-weight LLM families, revealing that closed-weight models with guideline-anchored prompting achieve robust performance within domains and language pairs.
- We demonstrate that a lightweight `ALOPE`-inspired approach, which leverages selected intermediate Transformer layer representations, achieves competitive QE performance in resource-constrained settings. Across the majority of the domains and language pairs, intermediate Transformer layers consistently yield stronger QE signals than final-layer representations.
- We establish a simple practical framework for QE deployment in low-resource, domain-sensitive settings, providing clear guidance on when to prioritise strong prompting versus when to apply lightweight adapter-based methods.

2 Background

Machine Translation (MT) quality remains uneven across language pairs and domains despite advances in neural and Transformer-based architectures (Specia et al., 2018; Zhao et al., 2024). This disparity is pronounced for English→Indic translation, where rich morphology, code-mixing, script diversity, and limited parallel data pose persistent challenges (Sindhujan et al., 2025b). Domain-specific translation in high-risk contexts (Healthcare, Legal) is particularly fragile, as publicly available corpora are skewed toward general web and news content, leaving specialised terminology and discourse structures under-represented (Specia et al., 2018; Zhao et al., 2024). Recent domain-focused evaluations further show that MT quality is strongly domain dependent: systems that perform well on General text often degrade substantially in specialised domains such as Healthcare, Legal, Literary, or User-generated content (Mäkelä et al., 2024; Gupta et al., 2024). In the Healthcare domain, LLMs leveraging document-level context have been shown to outperform traditional neural

MT systems, while this advantage largely disappears in General or News translation, highlighting the interaction between domain characteristics and model effectiveness (Mäkelä et al., 2024).

Quality Estimation (QE) addresses this by predicting translation quality without reference translations, enabling scalable assessment (Zerva et al., 2022). Segment-level QE using Direct Assessment (DA) scores provides a practical alternative to fine-grained frameworks like MQM, balancing interpretability with annotation efficiency (Graham et al., 2013). However, QE performance in low-resource, domain-shifted settings remains constrained by limited labelled data and high-impact error types (Zhao et al., 2024), which are particularly prevalent in specialised domains where terminology misuse or semantic distortion can be critical.

LLMs offer prompt-based, reference-free QE but suffer from score compression, inconsistent calibration, and weak sensitivity to subtle errors due to optimisation for next-token prediction rather than regression objectives (Zerva et al., 2024; Kočmi and Federmann, 2023). Evidence from domain adaptation studies suggests that while fine-tuning LLM-based MT models on in-domain data consistently improves translation quality, the effectiveness of adaptation depends on the availability and diversity of domain-specific data (Patel et al., 2024). In the `Legal` domain, fine-tuning a strong multilingual pre-trained model has been shown to outperform models explicitly pre-trained for the target domain, indicating that robust `General`-purpose representations combined with targeted adaptation are often more effective than domain-specific pre-training alone (Singh et al., 2025). When in-domain data is scarce or rapid adaptation is required, in-context learning provides a viable alternative: selecting few-shot examples based on topic similarity can significantly improve translation quality for unseen domains, though gains depend on balancing relevance with sufficient example diversity (Li et al., 2024).

Cross-lingual QE signals are not uniformly distributed across Transformer layers, intermediate layers encode more stable semantic and alignment information than final layers, especially for low-resource languages (Kargaran et al., 2025; Tenney et al., 2019). Parameter-efficient methods such as Low-Rank Adaptation (LoRA) enable task-specific fine-tuning by updating minimal parameters while keeping base models frozen (Hu et al., 2022; Dettmers et al., 2023). The `ALOPE` framework

Domain	**Langs**	**Train**	**Test**
`Healthcare`	Hi/Mr/Ta/Gu	13,280	1,660
`Legal`	Gu/Ta/Te	6,160	770
`Tourism`	Hi/Mr/Te	13,840	1,730
`General`	Hi/Mr/Ta/Te/Gu	18,880	2,360

Table 1: Indic-Domain-QE dataset composition, showing domain coverage, language pairs, and the number of instances in the train and test splits aggregated across languages.

(§ 3.3) extends this by attaching lightweight regression heads to intermediate layers, targeting QE-relevant representations (Sindhujan et al., 2025c). Our work investigates whether intermediate-layer adaptation provides consistent benefits across domains and low-resource English→Indic pairs, or whether domain characteristics and pre-training coverage moderate its effectiveness.

3 Methodology

The methodology adopted in this research provides a systematic and reproducible approach to investigating domain-specific QE for English–Indic language pairs using LLMs. Unlike generic MT evaluation, this research emphasises four practical domains where translation errors can have tangible effects on daily life. Starting from a structured dataset with human-annotated DA scores across four domains and five language pairs, we evaluate QE approaches along two parallel tracks: Prompt-Only approaches and the ALOPE-based approach. Within Prompt-only approaches, we evaluate closed-weight models for comparison, whereas open-weight models are used as backbone within both approaches. Figure 1 provides an overview of our methodological framework using open-weight models.

3.1 Dataset Construction

The Indic-Domain-QE dataset was created to support systematic evaluation of QE in domain-sensitive contexts. Texts were sourced from publicly available bilingual resources and curated domain-specific materials. Domain characterisation follows standard MT and QE research practices (Specia et al., 2018; Zhao et al., 2024), based on provenance and communicative function of source texts.

`Healthcare` data consist of patient-facing medical content such as information leaflets, public health advisories, and community health ar-

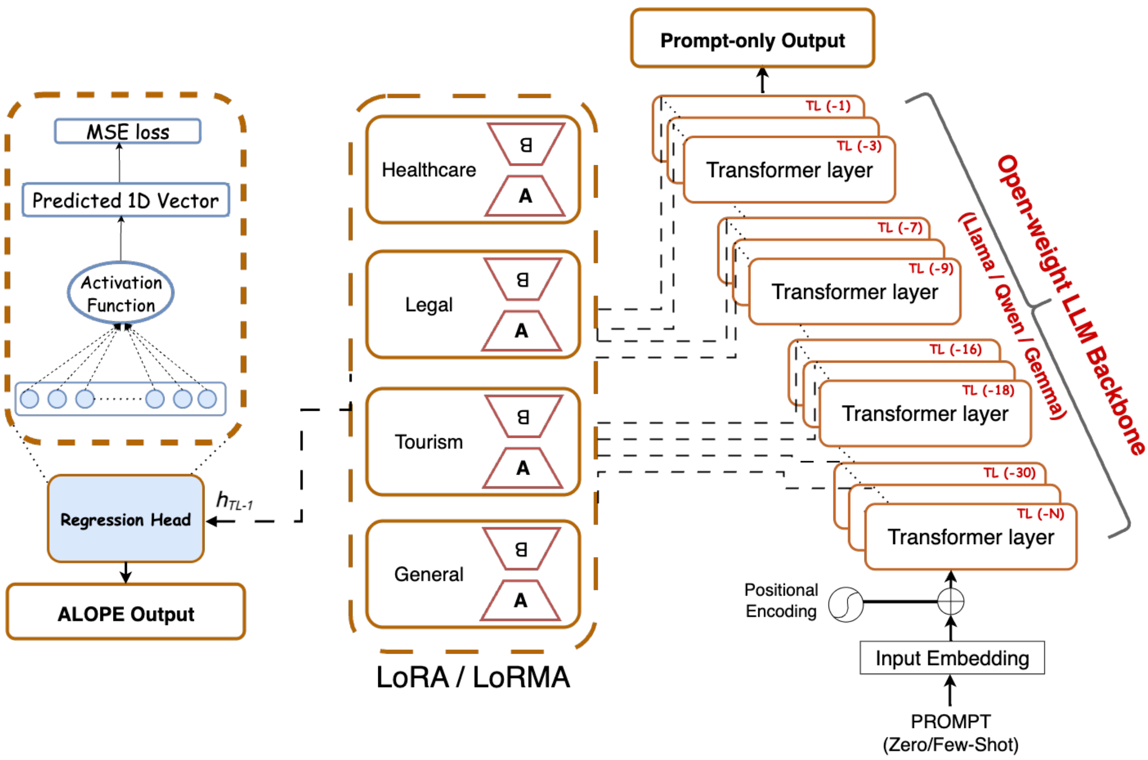

Figure 1: Methodological Framework uses open-weight models for (i) prompt-only approaches (zero-shot, few-shot, guideline-anchored), and (ii) ALOPE adaptation with LoRA/LoRMA.

ticles, exhibiting high terminology density and sensitivity to negation and numerical expressions. `Legal` texts are drawn from contracts, policy excerpts, and official notices, characterised by formal register, modality, and scope-defining constructions. `Tourism` data originate from brochures, attraction descriptions, and travel guidelines, rich in named entities and culturally grounded references. `General`-domain data comprise broad-coverage sentences from encyclopaedic and public-information sources, providing a baseline with minimal domain-specific constraints.

Each entry contains an English source, its translation, a domain label, and human-annotated DA scores. Annotators were trained following established DA protocols (Graham et al., 2013). Table 1 provides a detailed breakdown of the instances used in this study, with the training split employed for fine-tuning ALOPE and the test split used for evaluation under both prompt-only and ALOPE-based settings.

3.2 Prompt-only Approaches

We evaluate three prompting strategies for QE, while keeping the task instruction fixed across domains and language pairs, varying only the presence of in-context examples and explicit guidance. Selected closed-weight and open-weight models are evaluated under zero-shot prompting, few-shot prompting without guidelines, and few-shot prompting with guideline anchoring.

Zero-shot. The model receives only a natural-language task instruction and the input sentence pair, without in-context demonstrations (Prompt: App. A-Figure 4). This setup relies entirely on knowledge from pre-training of the selected model(Brown et al., 2020). While simple and cost-effective, zero-shot prompting often leads to score compression and unstable calibration in regression tasks.

Few-shot (without guidelines). This includes 1 to 5 labelled input-output examples in the prompt, with minimal additional instruction (Prompt: App. A - Figure 2). This in-context learning setup conditions the model on representative examples without parameter updates (Brown et al., 2020). Prior work shows few-shot exemplars improve task adherence but may suffer from variability when explicit scoring criteria are absent.

Few-shot (with guidelines). This augments the exemplar-based setup with an explicit scoring rubric defining the intended semantics of the output scale (Prompt: App. A-Figure 3). Guidelines clarify numerical score assignment, reducing ambiguity. Prior studies demonstrate that such explicit constraints improve output consistency and reduce prompt sensitivity (Mishra et al., 2022; Zhao et al., 2021).

Prompt-only evaluations are conducted with two types of models:

- **Closed-weight models:** Gemini-1.5-Pro and Gemini-2.5-Pro, accessed via API, serving as strong prompt-only baselines without parameter updates.
- **Open-weight models:** LLaMA-3.2-3B Instruct, LLaMA-3.1-8B Instruct, Qwen3-14B Instruct, and Gemma-3-27B Instruct.

3.3 ALOPE Framework

`ALOPE` (**A**daptive **L**ayer **OP**timization for Translation Quality **E**stimation) (Sindhujan et al., 2025c) is a parameter-efficient fine-tuning-based framework that attaches regression heads to selected intermediate Transformer layers and updates only a minimal parameter subset using LoRA (Hu et al., 2022). The original framework explores both single-layer heads and multi-layer variants with dynamic weighting. We only adopt the simplified single-layer configuration for feasibility and reproducibility, evaluating both LoRA and LoRMA adapter variants within the same framework. LoRMA stands for Low-Rank Multiplicative Adaptation, and unlike LoRA, which fine-tunes models by additively injecting low-rank weight updates, LoRMA adapts models by multiplicatively modulating existing weights (Bihany et al., 2025).

We conduct all `ALOPE` experiments using LLaMA-3.2-3B Instruct as the backbone model, following the best-performing configuration reported in the original `ALOPE` study (Sindhujan et al., 2025c). We explore rank configurations $R \in \{32, 64, 128\}$ with scaling factor $\alpha \in \{16, 32\}$, extracting representations from Layers $\{-1, -7, -9, -11\}$. The rank R determines the capacity of the low-rank decomposition used to parameterise the weight update, with each update factorised into two trainable matrices while keeping the pre-trained weights frozen. The scaling factor α rescales the update by α/R, stabilising training and enabling effective adaptation across different rank settings without extensive hyperparameter tuning (Hu et al., 2022).

The regression head is a lightweight two-layer feed-forward network with ReLU activation, mapping layer representations to scalar DA predictions. Training uses mean squared error (MSE) loss on gold DA scores. All the adapter-based experiments use 4-bit quantized base models via QLoRA (Quantized-LoRA) to ensure computational efficiency (Dettmers et al., 2023).

Evaluation Metrics. We evaluate model performance using two correlation-based metrics. Spearman's rank correlation (ρ) (Sedgwick, 2014) measures the agreement between the relative ordering of predicted and gold Direct Assessment (averaged across annotators) scores, making it well-suited for Quality Estimation where reliable ranking of translations is often more critical than exact score values. Owing to its robustness to scale differences and outliers, ρ is used as our primary evaluation metric. Pearson's correlation (r) (Cohen et al., 2009) quantifies linear agreement between predicted and gold DA scores, indicating how closely model outputs match human scores on the same numeric scale. We report r as the additional metric. Both metrics are computed per language pair and domain, and macro-averaged to obtain domain-level results.

4 Results

This section presents a comprehensive analysis of QE performance across domains, languages, prompting strategies, and parameter-efficient adaptation methods. We begin by analysing the adapter-based approaches using `ALOPE`, which highlights how Transformer layer selection affects QE performance. We then contextualise these findings through domain-wise comparisons against prompt-only baselines, enabling a detailed understanding of when and why lightweight adaptation becomes necessary.

4.1 Layer-wise Analysis with ALOPE

Table 2 shows that, for the `General` domain, the highest Spearman correlations for most language pairs are achieved at intermediate Transformer layers when using `ALOPE` with LoRA. The same pattern is observed consistently across `Healthcare`, `Legal`, and `Tourism` domains. This behaviour is mirrored by `ALOPE` with LoRMA (Table 3), where peak correlations for individual language pairs are

	Layer	$R = 32, \alpha = 16$						$R = 64, \alpha = 32$						$R = 128, \alpha = 32$					
		En-Hi	En-Mr	En-Ta	En-Te	En-Gu	Avg	En-Hi	En-Mr	En-Ta	En-Te	En-Gu	Avg	En-Hi	En-Mr	En-Ta	En-Te	En-Gu	Avg
General	L −1	0.363	-0.026	0.385	0.180	0.190	0.218	**0.476**	0.155	0.605	0.267	0.435	†0.388	0.384	0.105	0.524	0.196	0.391	0.320
	L −7	0.171	**0.157**	0.389	0.176	0.212	0.221	0.213	0.059	0.469	0.252	0.371	0.273	0.057	0.057	0.346	0.134	0.320	0.183
	L −9	0.123	0.090	0.531	0.163	0.458	0.273	0.311	0.038	0.531	0.279	**0.485**	0.329	0.051	0.107	0.300	0.061	0.343	0.172
	L −11	0.178	0.061	0.492	-0.029	0.484	0.237	0.321	0.034	0.568	-0.021	0.480	0.276	0.308	0.007	**0.610**	**0.292**	0.457	0.335
Healthcare	L −1	**0.520**	0.025	0.201	NA	0.377	0.281	0.382	0.033	0.248	NA	0.490	0.288	0.346	0.156	0.165	NA	0.452	0.280
	L −7	0.168	0.150	0.154	NA	0.346	0.204	0.314	0.021	0.363	NA	**0.532**	0.307	0.283	0.182	0.392	NA	0.477	†0.333
	L −9	0.110	0.096	0.086	NA	0.255	0.137	0.262	0.132	0.414	NA	0.417	0.306	0.161	0.019	0.321	NA	0.358	0.215
	L −11	0.123	0.050	0.352	NA	0.435	0.240	0.265	**0.192**	0.320	NA	0.469	0.311	0.346	-0.042	**0.415**	NA	0.511	0.308
Legal	L −1	NA	NA	-0.091	-0.009	0.002	-0.033	NA	NA	0.518	0.071	0.259	0.283	NA	NA	0.380	0.136	0.355	0.290
	L −7	NA	NA	0.379	0.052	0.367	0.266	NA	NA	0.443	0.059	0.366	0.289	NA	NA	0.478	0.124	0.280	0.294
	L −9	NA	NA	0.558	0.143	0.305	0.335	NA	NA	0.521	0.079	0.278	0.293	NA	NA	0.432	0.107	0.283	0.274
	L −11	NA	NA	0.489	0.043	0.243	0.258	NA	NA	0.580	**0.267**	**0.445**	0.430	NA	NA	**0.581**	0.267	0.445	†0.431
Tourism	L −1	-0.153	-0.009	NA	0.002	NA	-0.053	-0.080	0.029	NA	0.065	NA	0.047	-0.061	0.045	NA	0.080	NA	0.063
	L −7	-0.104	0.516	NA	0.019	NA	0.144	0.125	0.566	NA	0.154	NA	0.282	0.180	0.600	NA	0.210	NA	0.330
	L −9	-0.082	0.579	NA	-0.050	NA	0.149	0.288	0.596	NA	0.189	NA	0.357	0.330	0.640	NA	**0.220**	NA	0.397
	L −11	0.298	0.633	NA	0.183	NA	0.371	0.298	0.633	NA	0.183	NA	0.371	**0.350**	**0.670**	NA	0.205	NA	†0.408

Table 2: Spearman's (ρ) scores obtained for ALOPE layer-wise (L) experiments with LoRA across different domains and language pairs. The bolded values represent the highest spearman scores obtained for each language pair in each domain. The (†) represents the highest average obtained in each domain. 'NA' indicates that data from that specific domain and language pair is unavailable.

likewise concentrated at intermediate layers, particularly Layers -7, -9, and -11.

To summarise the overall trends, we visualise the domain-level average Spearman correlations from Tables 2 and 3 in Figures 6 and 7 in Appendix E. Each figure plots the average correlation computed across the five language pairs for a given domain—against the Transformer layer from which the regression head extracts representations. Across all domains, the averaged results reveal a clear and consistent pattern: intermediate Transformer layers, particularly Layers -9 and -11, yield substantially higher Spearman correlations than the final layer (Layer -1). This trend is stable across domains and adapter variants, and aligns with prior findings of ALOPE (Sindhujan et al., 2025c). These results support the hypothesis that QE-relevant signals are more robustly encoded in intermediate representations for English→Indic language pairs, while final-layer representations are more specialised for next-token prediction and instruction-following objectives.

Further analysis of the averaged correlations indicates that ALOPE with LoRMA introduces a stabilising effect across layers (Appendix E, Figure 7). Compared to LoRA, LoRMA produces smoother layer-wise behaviour, mitigating extremely low correlations at shallow layers (e.g., Layers -1 and -7) and reducing variance between adjacent layers across most configurations. This stabilisation is most pronounced in the General and Legal domains, where performance becomes less sensitive to the exact choice of layer.

Despite these stability gains, LoRMA obtains competitive performance to LoRA in the Tourism (0.408 vs. 0.404) domain. In Healthcare and Legal domains, LoRA consistently delivers higher correlations. These findings highlight a principled trade-off: LoRA is preferable when maximising ranking accuracy is the primary objective, whereas LoRMA offers increased robustness when deployment constraints limit precise layer selection.

Adapter configurations. ALOPE has different adapter configurations as explained in the section 3.3. Lower-rank adapters with $R = 32$ consistently underfit across several domains, achieving substantially lower correlations than higher-rank configurations. Increasing adapter capacity to $R = 128$ occasionally improves the highest correlations but introduces instability and higher variance, particularly for shallow layers. In contrast, $R = 64$ with $\alpha = 32$ consistently provides the best balance between expressive capacity and robustness across domains.

4.2 Prompt-only QE Baselines

Table 4 shows the best Spearman scores obtained with prompt-only baselines for each language pair and domain with closed- and open-weight models (Detailed model-wise results are reported in Appendix F, G, H).

Closed-weight models consistently provide strong performance across domains, even under zero-shot prompting. This behaviour, which yields competitive scores with minimal prompt engineering, can be attributed to the large scale of the Gemini models and their extensive pre-training. Open-weight models exhibit substantially weaker and more variable behaviour under prompt-only evaluation, particularly in `Healthcare` and `Legal`. Overall, considering both open- and closed-weights models, few-shot prompting improves performance, and guideline-based prompts further stabilise the Spearman scores.

4.3 ALOPE vs. Prompt-only Approaches

Since `ALOPE` was evaluated exclusively with LLaMA-3.2-3B, our comparative analysis focuses on prompt-only results obtained with the same backbone model (Appendix I). Across the majority of language pairs and domains, `ALOPE` consistently achieves higher Spearman correlations than prompt-only prompting strategies. These results demonstrate that `ALOPE` provides a practical and effective approach for improving QE, even when applied to smaller open-weight LLMs with quantization and lightweight adapters, which substantially reduce parameter count, model size, and computational cost compared to large closed-weight models. While `ALOPE` is applicable to other open-weight architectures, extending this analysis to additional models is left for future work.

4.4 Domain-specific Observations

This section analyses performance differences between the best prompt-only baselines (Table 4) and the `ALOPE`-based approaches (Tables 2 and 3). Across all settings, the `General` domain consistently achieves the highest correlations, reflecting its broader linguistic coverage and lower terminological complexity compared to specialised domains.

`Healthcare` presents a mixed picture for `ALOPE`'s effectiveness. While prompt-only baselines achieve strong correlations for several language pairs, `ALOPE` with intermediate-layer adaptation shows limited improvements and occasionally underperforms these baselines. Notable exceptions include English→Gujarati, where `ALOPE` achieves competitive performance. This suggests that `Healthcare` QE benefits more from strong prompting strategies with closed-weight models than from lightweight adapter-based methods for most language pairs. When `ALOPE` does improve performance, gains are strongest at the Transformer Layers -7 to -11 (Table 2) for this domain.

`Legal` remains the most challenging domain overall. `ALOPE` shows selective improvements, particularly for English→Tamil (ρ = 0.581, outperforming open-weight prompt baselines), though absolute correlations remain lower than other domains. Performance gains are configuration-sensitive, reflecting the strict semantic requirements of `Legal` text.

`Tourism` exhibits an irregular pattern where zero-shot prompting with open-weight models achieves surprisingly strong performance (average ρ = 0.646), often matching or exceeding average Spearman's obtained with stronger closed-weight models and `ALOPE`. For English→Marathi, `ALOPE` achieves ρ = 0.670, which is competitive but does not surpass the best closed-weight prompt-only baseline (few-shot with guideline: ρ = 0.702). Gains for other language pairs level off quickly, consistent with the entity-heavy and descriptive nature of `Tourism` content, where surface-level fluency and entity preservation may be adequately captured by prompting alone.

5 Discussion

Closed-weight models achieve strong and stable correlations even under zero-shot prompting, reflecting their extensive instruction tuning and multilingual pre-training. Guideline-anchored prompts further improve robustness by clarifying scale semantics and reducing mid-range score compression (Mishra et al., 2022; Zhao et al., 2021). As a result, closed-weight models represent the most reliable option when API access is available. In contrast, open-weight models exhibit substantially weaker and more variable performance under prompt-only

Domain	Layer	R = 64, α = 32						R = 128, α = 32					
		En-Hi	En-Mr	En-Ta	En-Te	En-Gu	Avg	En-Hi	En-Mr	En-Ta	En-Te	En-Gu	Avg
General	L −1	0.299	0.068	0.481	0.114	0.298	0.252	0.276	0.092	0.330	0.059	0.295	0.210
	L −7	0.324	**0.314**	0.438	0.087	0.381	0.309	**0.392**	0.081	**0.507**	0.155	**0.421**	0.311
	L −9	0.255	0.265	0.473	0.070	0.362	0.285	0.278	0.283	0.484	0.101	0.386	0.306
	L −11	0.337	0.267	0.483	**0.168**	0.391	†0.329	0.343	0.259	0.409	0.086	0.404	0.300
Healthcare	L −1	-0.316	-0.021	-0.381	NA	0.420	-0.075	-0.076	-0.025	0.374	NA	**0.512**	0.196
	L −7	0.262	0.124	0.405	NA	0.495	0.322	**0.365**	-0.017	**0.413**	NA	0.457	0.305
	L −9	0.077	0.074	0.400	NA	0.453	0.251	0.343	**0.131**	0.277	NA	0.490	0.310
	L −11	0.315	0.016	0.412	NA	0.482	0.306	0.351	0.105	0.392	NA	0.467	†0.329
Legal	L −1	NA	NA	0.509	0.073	0.182	0.255	NA	NA	**0.515**	0.104	0.173	0.264
	L −7	NA	NA	0.488	0.092	0.193	0.258	NA	NA	0.478	0.077	0.214	0.256
	L −9	NA	NA	0.488	0.112	**0.240**	†0.280	NA	NA	0.430	**0.125**	0.216	0.257
	L −11	NA	NA	0.424	0.058	0.178	0.220	NA	NA	0.480	0.111	0.110	0.234
Tourism	L −1	-0.315	0.462	NA	-0.082	NA	0.022	-0.354	-0.013	NA	-0.171	NA	-0.179
	L −7	0.436	0.493	NA	0.184	NA	0.371	0.446	0.502	NA	**0.197**	NA	0.382
	L −9	0.453	**0.532**	NA	0.227	NA	†0.404	**0.465**	0.445	NA	0.142	NA	0.351
	L −11	0.389	0.512	NA	0.074	NA	0.325	0.423	0.468	NA	0.164	NA	0.352

Table 3: Spearman's (ρ) scores obtained for ALOPE layer-wise (L) experiments with LoRMA across different domains and language pairs. The bolded values represent the highest spearman scores obtained for each language pair in each domain. The (†) represents the highest average obtained in each domain. 'NA' indicates that a language pair is unavailable for that specific domain.

evaluation, with zero-shot prompting often yielding near-zero or negative correlations. This highlights the inherent limitations of prompt engineering alone for smaller models, particularly in high-risk domains (Sindhujan et al., 2025b).

ALOPE provides a targeted remedy for this gap, but its effectiveness is strongly domain-dependent. In the Legal domain, ALOPE substantially improves performance, underscoring the importance of domain-specific adaptation for semantically precise content. In contrast, gains in Healthcare are limited, suggesting that this domain benefits more from broad pre-training coverage of medical terminology in large closed-weight models than from lightweight adapter-based fine-tuning. General and Tourism domains exhibit intermediate behaviour, indicating that the utility of ALOPE depends on the interaction between domain complexity and pre-training corpus characteristics. Incorporating LoRMA within ALOPE further introduces stability-oriented regularisation, yielding smoother layer-wise behaviour and mitigating lower correlations at shallow layers (Sindhujan et al., 2025c).

The layer-wise superiority of intermediate representations (Layers -9, -11) holds consistently across all five languages, suggesting this reflects fundamental properties of how multilingual LLMs encode cross-lingual semantic alignment rather than language-specific artifacts (Kargaran et al., 2025). This consistency across typologically diverse languages strengthens the generalisability of intermediate-layer adaptation for QE in low-resource settings.

Taken together, our results motivate a conditional deployment strategy. When access to closed-weight models is feasible, guideline-anchored prompting offers the most reliable solution. When deployment constraints such as cost, latency, or privacy preclude such access, ALOPE with LoRA provides a lightweight and effective alternative, particularly for semantically complex domains such as Legal. This underscores the importance of empir-

Domain	Prompt setting	Model family	En-Hi	En-Mr	En-Ta	En-Te	En-Gu	Average
`General`	Zero-shot	Closed	0.424	**0.597**	0.848	0.392	**0.924**	†0.637
		Open	0.390	-0.058	0.772	0.382	0.812	0.460
	Few-shot + Guidelines	Closed	0.475	0.582	0.886	0.238	0.776	0.591
		Open	0.408	0.038	0.832	0.442	0.849	0.514
	Few-shot (No Guidelines)	Closed	**0.563**	0.314	**0.940**	-0.031	0.860	0.529
		Open	0.418	0.375	0.867	**0.486**	0.752	0.580
`Healthcare`	Zero-shot	Closed	0.126	0.814	0.366	NA	0.346	0.413
		Open	0.569	0.389	**0.603**	NA	**0.494**	0.514
	Few-shot + Guidelines	Closed	0.415	0.669	-0.040	NA	0.100	0.286
		Open	0.447	**0.884**	0.411	NA	0.211	0.488
	Few-shot (No Guidelines)	Closed	0.585	0.786	0.168	NA	0.458	0.385
		Open	**0.611**	**0.884**	0.422	NA	0.398	†0.579
`Legal`	Zero-shot	Closed	NA	NA	**0.749**	0.109	0.677	0.512
		Open	NA	NA	0.418	0.287	0.265	0.323
	Few-shot + Guidelines	Closed	NA	NA	0.717	0.530	0.699	†0.649
		Open	NA	NA	0.418	0.528	**0.727**	0.558
	Few-shot (No Guidelines)	Closed	NA	NA	0.475	0.230	0.475	0.393
		Open	NA	NA	0.737	**0.687**	0.473	0.632
`Tourism`	Zero-shot	Closed	0.416	0.474	NA	0.217	NA	0.369
		Open	0.613	0.689	NA	**0.636**	NA	†0.646
	Few-shot + Guidelines	Closed	0.502	0.679	NA	0.472	NA	0.551
		Open	0.158	**0.702**	NA	0.583	NA	0.481
	Few-shot (No Guidelines)	Closed	0.509	0.685	NA	0.397	NA	0.530
		Open	**0.737**	0.687	NA	0.473	NA	0.632

Table 4: Reports best Spearman's correlation (ρ) achieved for each domain and language pair across all evaluated models, including both open- and closed-weight LLMs. The bolded values represent the highest spearman scores obtained for each language pair in each domain. The (†) represents the highest average obtained in each domain. Detailed model-wise results are provided in Appendices F, G, and H.

ical validation prior to adopting adaptation strategies, as their effectiveness depends on domain-specific interactions with pre-training data.

6 Conclusion

This work investigates domain-specific quality estimation for English→Indic translation across `Healthcare`, `Legal`, `Tourism`, and `General` domains, covering five language pairs (Hindi, Marathi, Tamil, Telugu, Gujarati). We systematically compare prompt-only approaches with parameter-efficient `ALOPE` adaptation, revealing that closed-weight models with guideline-anchored prompting achieve robust performance without parameter updates, while open-weight models exhibit substantial fragility under prompt-only evaluation. `ALOPE` demonstrates reasonable QE performance even with smaller open-source LLMs when intermediate Transformer layer embeddings are utilised for quality estimation. Our findings support a conditional deployment strategy: prioritise closed-weight prompting when API access is viable; apply `ALOPE` with LoRA for open-weight models in resource-constrained environment; and use `ALOPE` with LoRMA when precise layer tuning is constrained. Future work should investigate multi-layer fusion approaches and interpretability techniques to understand which linguistic phenomena drive layer-specific improvements.

7 Limitations

Our study is limited to English→Indic language pairs across four domains, due to the limited availability of QE data for other language pairs and domains. Further work is needed to examine the generalisability of these findings to other language families and more specialised technical domains. In addition, all `ALOPE` fine-tuning experiments use a relatively small backbone model (LLaMA-3.2-3B Instruct) due to computational constraints, and results may differ when scaling to larger open-weight models.

References

Harsh Bihany, Shubham Patel, and Ashutosh Modi. 2025. LoRMA: Low-rank multiplicative adaptation for LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10113–10133, Vienna, Austria. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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

Tim Dettmers and 1 others. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint*, arXiv:2305.14314.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. *Proceedings of the 7th Linguistic Annotation Workshop*.

Deepak Gupta, Rishabh Agarwal, and Swaroop Mishra. 2024. Domain dynamics: Evaluating large language models in english-hindi translation. In *Proceedings of the International Conference on Computational Linguistics (COLING 2024)*. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phil Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. *International Conference on Learning Representations (ICLR)*.

Amir Kargaran and 1 others. 2025. Layer-wise cross-lingual alignment in low-resource languages. *arXiv preprint*.

Tom Kočmi and Christian Federmann. 2023. Gemba-mqm: Detecting translation quality error spans with gpt-4. In *Proceedings of the Eighth Conference on Machine Translation*. Association for Computational Linguistics.

Yifan Li, Yimeng Chen, and Min Zhang. 2024. Topic-guided example selection for domain adaptation in LLM-based machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Arle Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2013. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.

Eetu Mäkelä, Saad Khalid, Lorenzo Gatti, and Matteo Negri. 2024. OpenWHO: A document-level parallel corpus for health translation in low-resource languages. In *Proceedings of the Ninth Conference on Machine Translation (WMT 2024)*. Association for Computational Linguistics.

Swaroop Mishra and 1 others. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

Shubham Patel, Harsh Bihany, and Ashutosh Modi. 2024. Exploring the effectiveness of LLM domain adaptation for business IT machine translation. In *Proceedings of the Workshop on Domain Adaptation and Representation Transfer*. Association for Computational Linguistics.

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

Archchana Sindhujan, Diptesh Kanojia, Constantin Orasan, and Shenbin Qian. 2025a. When LLMs struggle: Reference-less translation evaluation for low-resource languages. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 437–459, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Archchana Sindhujan, Diptesh Kanojia, and Constantin Orăsan. 2025b. Reference-less evaluation of machine translation: Navigating through the resource-scarce scenarios. *Information*, 16(10).

Archchana Sindhujan, Shenbin Qian, Chan Chi Chun Matthew, Constantin Orăsan, and Diptesh Kanojia. 2025c. ALOPE: Adaptive layer optimization for translation quality estimation using large language models. *arXiv preprint*, arXiv:2508.07484.

Aayush Singh, Meet Patel, Ravi Shah, and Pushpak Bhattacharyya. 2025. Team-SVNIT at JUST-NLP 2025: Domain-adaptive fine-tuning of multilingual models for english-hindi legal machine translation. In *Proceedings of the Joint Workshop on NLP for Legal Applications*. Association for Computational Linguistics.

Lucia Specia, Dilip Raj, and 1 others. 2018. Quality estimation for machine translation. In *Machine Translation Summit XVI*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. *Proceedings of ACL*.

Chrysoula Zerva, Frédéric Blain, José G. C. De Souza, Diptesh Kanojia, Sourabh Deoghare, Nuno M. Guerreiro, Giuseppe Attanasio, Ricardo Rei, Constantin Orăsan, Matteo Negri, Marco Turchi, Rajen Chatterjee, Pushpak Bhattacharyya, Markus Freitag, and André Martins. 2024. Findings of the quality estimation shared task at WMT 2024: Are LLMs closing the gap in QE? In *Proceedings of the Ninth Conference on Machine Translation (WMT 2024)*, pages 82–109. Association for Computational Linguistics.

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

Haofei Zhao, Yilun Liu, Shimin Tao, Weibin Meng, Yimeng Chen, Xiang Geng, Chang Su, Min Zhang, and Hao Yang. 2024. From handcrafted features to llms: A brief survey for machine translation quality estimation. *arXiv preprint*, arXiv:2403.14118.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the International Conference on Machine Learning (ICML)*.

A Prompt Templates for Prompt-only Approaches

```
You are an expert in evaluating general-domain machine-translated content.
Rate the quality from 0 to 100 based on how accurately the MT output conveys the same meaning as the English source.

Only return the numeric score. No explanation.

Example 1:

Source (English): "The weather is pleasant today."

MT output (Hindi): "आज मौसम सुखद है।"

Score: 95


Example 2:

Source (English): "He went to the market to buy vegetables."

MT output (Marathi): "तो भाजीपाला आणण्यासाठी बाजारात गेला."

Score: 92


Example 3:
.....
Example 4:
.....
Example 5:
.....


Now rate the following:

Source (English): "{source}"

MT output ({target_lang}): "{mt_output}"

Score:
```

Figure 2: Few-shot QE Prompt (Without Guidelines)

You are an expert in evaluating machine-translated content.

Assess how accurately the MT output conveys the same meaning as the English source in a general-domain context.

Rate the quality from 0 to 100, where:

- 0–30: Very Poor (completely incorrect or misleading)

- 31–50: Poor (several errors, meaning not fully conveyed)

- 51–70: Moderate (some errors, but main idea understandable)

- 71–90: Good (minor issues but mostly accurate)

- 91–100: Excellent (highly accurate, fluent, and faithful)

Only return the numeric score. No explanation.

Example 1 (Very Poor – DA 22):

Source (English): "The driver was fined for jumping the red light."

MT output (Hindi): "ड्राइवर को इनाम मिला क्योंकि उसने लाल बत्ती पार की।"

Score: 22

Example 2 (Poor – DA 47):

Source (English): "The festival begins on the first full moon of spring."

MT output (Gujarati): "ઉત્સવ વસંતની શરૂઆતના દિવસે શરૂ થાય છે."

Score: 47

Example 3 (Moderate – DA 62):

.....

Example 4 (Good – DA 85):

.....

Example 5 (Excellent – DA 97):

.....

Now rate the following:

Source (English): "{source}"

MT output ({target_lang}): "{mt_output}"

Score:

Figure 3: Few-shot QE Prompt (With Guidelines)

```
You are an expert in evaluating machine-translated content. Assess how accurately the
MT output conveys the same meaning as the English source sentence. Rate the translation
quality on a scale from 0 to 100. Only return the numeric score. Do not provide any
explanation.

Source (English): "{source}"
MT output ({target_lang}): "{mt_output}"

Score:
```

Figure 4: Zero-shot prompt

B Prompt Templates for ALOPE

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

{source_lang} source: "{source_seg}"

{target_lang} translation: "{target_seg}"

Score:
```

Figure 5: ALOPE prompt

C Pearson's Correlation Scores obtained for ALOPE with LoRA

	Layer	$R = 32, \alpha = 16$						$R = 64, \alpha = 32$						$R = 128, \alpha = 32$					
		En-Hi	**En-Mr**	**En-Ta**	**En-Te**	**En-Gu**	**Avg**	**En-Hi**	**En-Mr**	**En-Ta**	**En-Te**	**En-Gu**	**Avg**	**En-Hi**	**En-Mr**	**En-Ta**	**En-Te**	**En-Gu**	**Avg**
General	L −1	0.400	-0.025	0.305	0.092	0.221	0.199	0.628	0.154	0.680	0.229	0.538	0.446	0.597	0.111	0.679	0.181	0.594	0.433
	L −7	0.538	0.117	0.645	0.244	0.272	0.363	0.488	0.053	0.674	0.216	0.562	0.399	0.306	0.043	0.527	0.177	0.471	0.305
	L −9	0.492	0.081	0.677	0.250	0.679	0.436	0.621	0.077	0.716	0.305	0.663	0.476	0.328	0.095	0.549	0.131	0.511	0.323
	L −11	0.539	0.080	0.720	0.016	0.687	0.408	0.582	0.019	0.674	0.031	0.637	0.389	0.585	0.039	0.647	0.257	0.583	0.422
Healthcare	L −1	0.589	0.001	0.405	NA	0.354	0.337	0.635	0.051	0.494	NA	0.440	0.405	0.707	0.164	0.464	NA	0.426	0.440
	L −7	0.514	0.158	0.502	NA	0.381	0.389	0.682	0.008	0.523	NA	0.499	0.428	0.708	0.180	0.617	NA	0.465	0.493
	L −9	0.554	0.076	0.435	NA	0.342	0.352	0.688	0.094	0.629	NA	0.405	0.454	0.616	0.002	0.608	NA	0.358	0.396
	L −11	0.625	0.015	0.611	NA	0.425	0.419	0.614	0.193	0.630	NA	0.469	0.477	0.685	-0.018	0.602	NA	0.496	0.441
Legal	L −1	NA	NA	0.059	0.002	0.011	0.024	NA	NA	0.468	0.040	0.318	0.276	NA	NA	0.430	0.100	0.382	0.304
	L −7	NA	NA	0.502	0.077	0.457	0.345	NA	NA	0.496	0.088	0.421	0.335	NA	NA	0.460	0.093	0.355	0.303
	L −9	NA	NA	0.551	0.142	0.387	0.360	NA	NA	0.459	0.059	0.334	0.284	NA	NA	0.452	0.081	0.354	0.295
	L −11	NA	NA	0.447	0.001	0.334	0.261	NA	NA	0.621	0.215	0.536	0.457	NA	NA	0.621	0.215	0.536	0.457
Tourism	L −1	0.059	0.002	NA	0.011	NA	0.024	0.119	0.064	NA	0.083	NA	0.089	0.130	0.075	NA	0.095	NA	0.100
	L −7	0.061	0.467	NA	0.145	NA	0.224	0.220	0.544	NA	0.286	NA	0.350	0.260	0.580	NA	0.320	NA	0.387
	L −9	0.103	0.557	NA	0.069	NA	0.243	0.317	0.614	NA	0.235	NA	0.389	0.360	0.660	NA	0.270	NA	0.430
	L −11	0.485	0.641	NA	0.219	NA	0.448	0.485	0.641	NA	0.219	NA	0.448	0.525	0.685	NA	0.245	NA	0.485

Table 5: ALOPE (LoRA) layer-wise (L) Pearson's (r) scores obtained across domains and language pairs for different adapter configurations.

D Pearson's Correlation Scores obtained for ALOPE with LoRMA

	Layer	$R = 64, \alpha = 32$						$R = 128, \alpha = 32$					
		En-Hi	**En-Mr**	**En-Ta**	**En-Te**	**En-Gu**	**Avg**	**En-Hi**	**En-Mr**	**En-Ta**	**En-Te**	**En-Gu**	**Avg**
General	L −1	0.262	0.030	0.552	0.053	0.430	0.259	0.217	0.120	0.383	-0.007	0.311	0.205
	L −7	0.382	0.315	0.557	0.067	0.469	0.358	0.407	0.083	0.513	0.093	0.472	0.314
	L −9	0.331	0.274	0.600	0.045	0.465	0.343	0.328	0.283	0.590	0.046	0.495	0.348
	L −11	0.394	0.272	0.598	0.105	0.491	0.372	0.395	0.271	0.540	0.052	0.492	0.350
Healthcare	L −1	0.420	0.008	-0.245	NA	0.391	0.144	0.198	0.009	0.306	NA	0.471	0.246
	L −7	0.224	0.130	0.300	NA	0.431	0.271	0.273	0.016	0.333	NA	0.429	0.263
	L −9	0.345	0.071	0.277	NA	0.411	0.276	0.414	0.136	0.285	NA	0.470	0.326
	L −11	0.374	0.034	0.335	NA	0.445	0.300	0.455	0.116	0.325	NA	0.424	0.330
Legal	L −1	NA	NA	0.375	0.032	0.304	0.237	NA	NA	0.479	0.100	0.339	0.306
	L −7	NA	NA	0.462	0.043	0.336	0.280	NA	NA	0.456	0.046	0.345	0.282
	L −9	NA	NA	0.467	0.113	0.365	0.315	NA	NA	0.464	0.139	0.356	0.323
	L −11	NA	NA	0.440	0.095	0.336	0.290	NA	NA	0.447	0.092	0.321	0.287
Tourism	L −1	-0.105	0.468	NA	0.176	NA	0.180	-0.283	-0.002	NA	0.134	NA	0.045
	L −7	0.384	0.492	NA	0.100	NA	0.325	0.396	0.512	NA	0.052	NA	0.320
	L −9	0.393	0.539	NA	0.153	NA	0.362	0.423	0.447	NA	0.089	NA	0.320
	L −11	0.360	0.524	NA	-0.057	NA	0.276	0.382	0.478	NA	0.043	NA	0.301

Table 6: ALOPE (LoRMA) layer-wise (L) Pearson's (r) scores obtained across domains and language pairs for different adapter configurations.

E ALOPE: Domain-wise Comparison of Average Performance

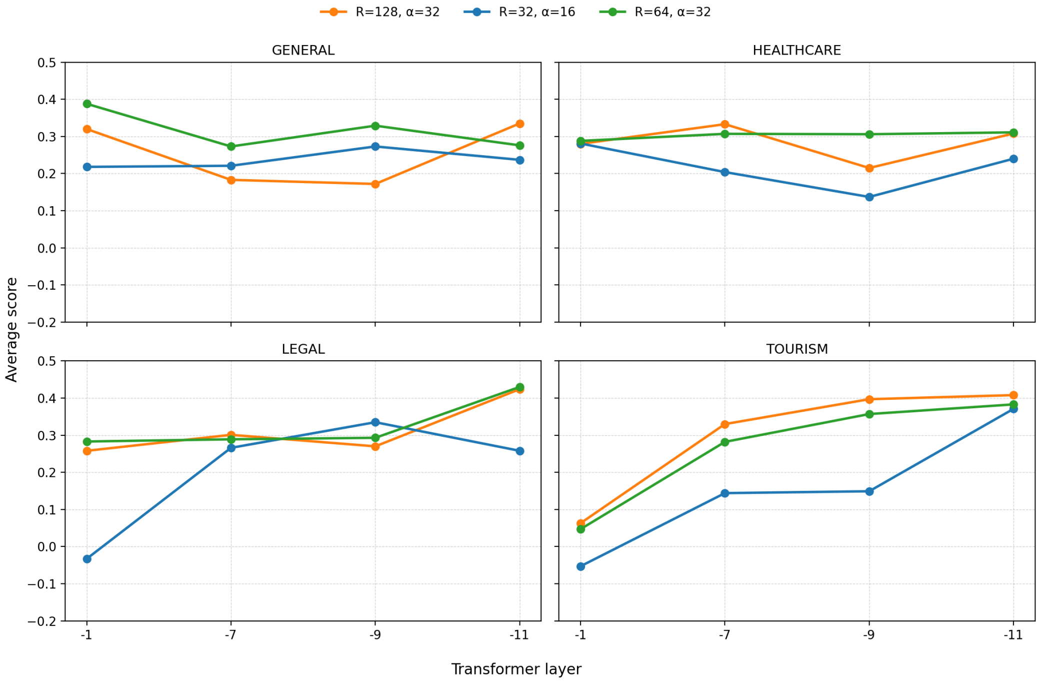

Figure 6: ALOPE with LoRA: Average Spearman's (ρ) across domains.

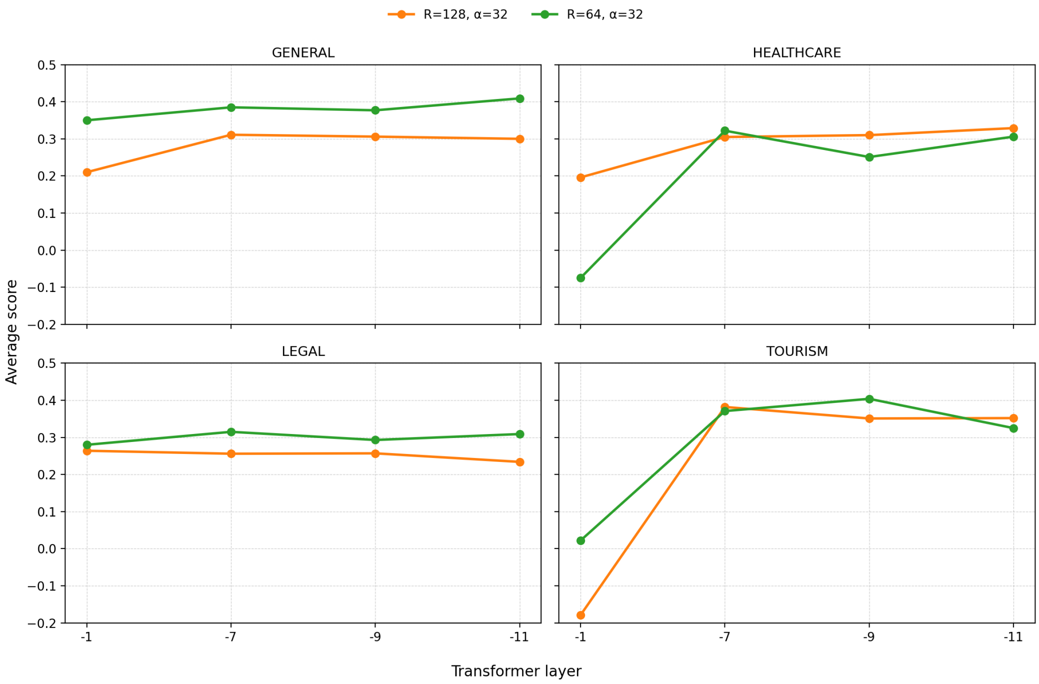

Figure 7: ALOPE with LoRMA: Average Spearman's (ρ) across domains.

F Zero-shot Evaluation Results

Domain	Model	En-Hi		En-Mr		En-Ta		En-Te		En-Gu		Avg	
		r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
General	Gemini-1.5-Pro	-0.082	-0.232	0.120	0.159	0.856	**0.848**	0.191	0.073	0.774	**0.924**	0.372	0.354
	Gemini-2.5-Pro	0.328	**0.424**	0.747	**0.597**	0.715	0.805	0.542	**0.392**	0.918	0.786	0.650	0.601
	Qwen3-14B	0.319	0.390	-0.159	-0.058	0.730	0.772	0.393	0.328	0.499	0.649	0.356	0.416
	Gemma3-27B	0.079	0.116	-0.259	-0.290	0.776	0.701	0.532	0.382	0.195	0.102	0.265	0.202
	LLaMA-3.2-3B	0.065	0.056	-0.184	-0.164	0.099	0.249	-0.366	-0.381	-0.174	-0.024	-0.112	-0.053
	LLaMA-3.1-8B	-0.156	-0.284	-0.512	-0.465	0.587	0.437	0.265	0.115	0.662	0.812	0.169	0.123
Healthcare	Gemini-1.5-Pro	0.134	0.039	0.668	**0.814**	-0.068	0.056	NA	NA	0.369	0.346	0.276	0.294
	Gemini-2.5-Pro	0.532	0.126	0.784	0.811	0.315	0.366	NA	NA	0.345	0.238	0.494	0.385
	Qwen3-14B	0.113	0.435	0.548	0.311	0.145	0.234	NA	NA	0.134	0.130	0.235	0.278
	Gemma3-27B	0.561	**0.569**	0.292	0.389	0.424	**0.603**	NA	NA	0.523	**0.494**	0.450	0.514
	LLaMA-3.2-3B	-0.462	-0.234	0.193	0.174	0.385	0.406	NA	NA	0.228	-0.055	0.086	0.073
	LLaMA-3.1-8B	0.001	-0.071	-0.196	-0.087	-0.180	0.100	NA	NA	-0.143	-0.395	-0.13	-0.113
Legal	Gemini-1.5-Pro	NA	NA	NA	NA	0.899	**0.749**	0.241	0.109	0.683	**0.677**	0.608	0.512
	Gemini-2.5-Pro	NA	NA	NA	NA	0.353	0.475	0.043	0.055	0.607	0.457	0.334	0.329
	Qwen3-14B	NA	NA	NA	NA	0.343	0.418	-0.689	-0.676	0.240	0.265	-0.035	0.002
	Gemma3-27B	NA	NA	NA	NA	-0.206	-0.056	-0.035	-0.045	0.260	0.201	0.006	0.033
	LLaMA-3.2-3B	NA	NA	NA	NA	-0.254	-0.174	-0.106	-0.011	-0.279	-0.429	-0.213	-0.205
	LLaMA-3.1-8B	NA	NA	NA	NA	0.013	0.031	0.270	**0.287**	0.045	-0.105	0.109	0.071
Tourism	Gemini-1.5-Pro	0.137	0.241	0.324	0.474	NA	NA	0.109	0.180	NA	NA	0.190	0.298
	Gemini-2.5-Pro	0.327	0.416	0.360	0.301	NA	NA	0.367	0.217	NA	NA	0.351	0.311
	Qwen3-14B	0.111	0.076	0.076	-0.058	NA	NA	0.588	0.541	NA	NA	0.258	0.186
	Gemma3-27B	0.129	0.102	0.804	0.654	NA	NA	0.633	**0.636**	NA	NA	0.522	0.464
	LLaMA-3.2-3B	0.468	0.417	0.432	0.469	NA	NA	0.459	0.309	NA	NA	0.453	0.398
	LLaMA-3.1-8B	0.582	**0.613**	0.839	**0.689**	NA	NA	0.734	0.584	NA	NA	0.718	**0.629**

Table 7: Zero-shot prompt-only QE performance. Spearman's (ρ) and Pearson's (r) scores are reported for all language pairs. Best ρ per language pair in **bold**.'NA' indicates that a language pair is unavailable for that specific domain.

G Few-shot with Guidelines Evaluation Results

Domain	Model	En-Hi		En-Mr		En-Ta		En-Te		En-Gu		Avg	
		r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
General	Gemini-1.5-Pro	-0.004	-0.094	-0.063	0.087	0.881	0.850	-0.319	-0.169	0.859	0.761	0.271	0.287
	Gemini-2.5-Pro	0.625	**0.475**	0.638	**0.582**	0.834	**0.886**	0.255	0.238	0.926	0.776	0.656	0.591
	Qwen3-14B	0.247	0.279	-0.159	-0.058	0.720	0.783	0.498	**0.442**	0.850	**0.849**	0.431	0.459
	Gemma3-27B	0.489	0.376	0.041	0.038	0.770	0.832	0.502	0.352	0.799	0.649	0.520	0.449
	LLaMA-3.2-3B	-0.637	-0.527	-0.101	-0.251	0.479	0.329	0.150	0.186	-0.332	-0.200	-0.088	-0.093
	LLaMA-3.1-8B	0.271	0.408	-0.116	0.034	-0.312	-0.339	-0.149	-0.145	-0.138	-0.288	-0.089	-0.066
Healthcare	Gemini-1.5-Pro	0.170	0.415	0.614	0.669	-0.237	-0.040	NA	NA	0.604	0.100	0.288	0.286
	Gemini-2.5-Pro	0.191	0.309	0.221	0.024	-0.302	-0.089	NA	NA	0.584	0.084	0.174	0.082
	Qwen3-14B	0.176	0.225	0.886	**0.884**	-0.144	-0.240	NA	NA	-0.357	-0.389	0.140	0.120
	Gemma3-27B	0.255	0.345	0.193	0.174	0.076	0.123	NA	NA	0.225	0.109	0.187	0.188
	LLaMA-3.2-3B	0.012	0.058	-0.499	-0.285	-0.476	-0.284	NA	NA	-0.381	-0.479	-0.336	-0.248
	LLaMA-3.1-8B	0.204	**0.447**	-0.541	-0.458	0.257	**0.411**	NA	NA	0.612	**0.211**	0.133	0.153
Legal	Gemini-1.5-Pro	NA	NA	NA	NA	0.867	**0.717**	0.446	**0.530**	-0.078	0.072	0.412	0.440
	Gemini-2.5-Pro	NA	NA	NA	NA	0.615	0.532	0.375	0.324	0.849	0.699	0.613	0.518
	Qwen3-14B	NA	NA	NA	NA	0.343	0.418	-0.689	-0.676	0.240	0.265	-0.035	0.002
	Gemma3-27B	NA	NA	NA	NA	0.249	0.099	0.195	0.102	0.699	**0.727**	0.381	0.309
	LLaMA-3.2-3B	NA	NA	NA	NA	-0.333	-0.477	0.378	0.528	0.293	0.304	0.113	0.118
	LLaMA-3.1-8B	NA	NA	NA	NA	0.254	0.104	0.121	0.030	-0.063	0.087	0.104	0.074
Tourism	Gemini-1.5-Pro	0.447	**0.502**	0.630	0.480	NA	NA	0.174	0.308	NA	NA	0.417	0.430
	Gemini-2.5-Pro	0.379	0.385	0.755	0.679	NA	NA	0.327	0.472	NA	NA	0.487	0.512
	Qwen3-14B	0.059	0.060	0.371	0.236	NA	NA	0.433	**0.583**	NA	NA	0.288	0.293
	Gemma3-27B	0.226	0.078	0.852	**0.702**	NA	NA	0.002	0.051	NA	NA	0.360	0.277
	LLaMA-3.2-3B	-0.471	-0.459	-0.248	-0.098	NA	NA	-0.118	0.032	NA	NA	-0.279	-0.175
	LLaMA-3.1-8B	0.008	0.158	-0.208	-0.058	NA	NA	-0.068	-0.218	NA	NA	-0.089	-0.039

Table 8: Few-shot with guidelines QE performance. Spearman's (ρ) and Pearson's (r) scores are reported for all language pairs. Best ρ per language pair in **bold**. 'NA' indicates that a language pair is unavailable for that specific domain.

H Few-shot without Guidelines Evaluation Results

Domain	Model	En-Hi		En-Mr		En-Ta		En-Te		En-Gu		Avg	
		r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
General	Gemini-1.5-Pro	0.107	0.022	-0.213	-0.103	0.744	0.743	-0.035	-0.185	0.941	0.791	0.309	0.254
	Gemini-2.5-Pro	0.713	**0.563**	0.447	0.314	0.888	**0.940**	-0.095	-0.031	0.939	**0.860**	0.578	0.529
	Qwen3-14B	-0.128	-0.040	-0.235	-0.234	0.761	0.867	0.319	0.235	0.658	0.752	0.275	0.316
	Gemma3-27B	0.177	0.234	0.489	**0.375**	0.519	0.450	0.636	**0.486**	0.765	0.627	0.517	0.434
	LLaMA-3.2-3B	0.497	0.418	-0.101	0.014	0.393	0.477	0.223	0.188	0.078	0.228	0.218	0.265
	LLaMA-3.1-8B	0.234	0.084	0.016	0.109	-0.332	-0.482	-0.180	-0.177	0.097	0.005	-0.033	-0.092
Healthcare	Gemini-1.5-Pro	0.568	0.307	0.852	0.552	-0.027	0.168	NA	NA	0.608	**0.458**	0.500	0.371
	Gemini-2.5-Pro	0.670	0.585	0.712	0.786	0.060	-0.086	NA	NA	0.388	0.393	0.458	0.420
	Qwen3-14B	0.261	0.413	0.865	**0.884**	-0.117	-0.075	NA	NA	-0.049	-0.143	0.240	0.270
	Gemma3-27B	0.083	0.187	0.097	0.114	0.132	0.275	NA	NA	0.216	0.174	0.132	0.188
	LLaMA-3.2-3B	0.513	**0.611**	0.123	0.035	0.648	**0.422**	NA	NA	0.139	0.071	0.356	0.285
	LLaMA-3.1-8B	0.519	0.406	-0.244	0.289	-0.101	0.250	NA	NA	0.283	0.398	0.114	0.336
Legal	Gemini-1.5-Pro	NA	NA	NA	NA	0.571	0.475	0.355	0.230	0.170	0.061	0.365	0.255
	Gemini-2.5-Pro	NA	NA	NA	NA	-0.342	-0.492	-0.262	-0.112	0.625	**0.475**	0.007	-0.043
	Qwen3-14B	NA	NA	NA	NA	0.350	0.467	0.365	0.215	0.332	0.473	0.349	0.385
	Gemma3-27B	NA	NA	NA	NA	0.277	0.311	0.828	**0.687**	0.304	0.211	0.470	0.403
	LLaMA-3.2-3B	NA	NA	NA	NA	-0.471	-0.459	-0.248	-0.098	-0.118	0.032	-0.279	-0.175
	LLaMA-3.1-8B	NA	NA	NA	NA	0.887	**0.737**	-0.438	-0.335	0.094	-0.026	0.181	0.125
Tourism	Gemini-1.5-Pro	0.142	0.193	0.835	0.685	NA	NA	-0.126	-0.208	NA	NA	0.284	0.223
	Gemini-2.5-Pro	0.359	0.509	0.501	0.406	NA	NA	0.422	0.397	NA	NA	0.427	0.437
	Qwen3-14B	0.350	0.467	0.365	0.215	NA	NA	0.332	**0.473**	NA	NA	0.349	0.485
	Gemma3-27B	0.277	0.311	0.828	**0.687**	NA	NA	0.304	0.211	NA	NA	0.470	0.403
	LLaMA-3.2-3B	-0.471	-0.459	-0.248	-0.098	NA	NA	-0.118	0.032	NA	NA	-0.279	0.175
	LLaMA-3.1-8B	0.887	**0.737**	-0.438	-0.335	NA	NA	0.094	-0.026	NA	NA	0.181	0.125

Table 9: Few-shot without guidelines QE performance. Spearman's (ρ) and Pearson's (r) scores are reported for all language pairs. Best ρ per language pair in **bold**. 'NA' indicates that a language pair is unavailable for that specific domain.

I Comparison of ALOPE and Prompt-only Approaches

Domain	Prompt setting	En-Hi	En-Mr	En-Ta	En-Te	En-Gu	Avg
General	Zero-shot	0.056	-0.164	0.249	-0.381	-0.024	-0.053
	Few-shot + Guidelines	-0.527	-0.251	0.329	0.186	-0.200	-0.093
	Few-shot (No Guidelines)	0.418	0.014	0.477	0.188	0.228	0.265
	ALOPE (LoRA)	**0.476**	0.157	**0.610**	**0.292**	**0.485**	†0.404
	ALOPE (LoRMA)	0.392	**0.314**	0.507	0.168	0.421	0.360
Healthcare	Zero-shot	-0.234	0.174	0.406	NA	-0.055	0.073
	Few-shot + Guidelines	0.058	-0.285	-0.284	NA	-0.479	-0.248
	Few-shot (No Guidelines)	**0.611**	0.035	0.422	NA	0.071	0.285
	ALOPE (LoRA)	0.520	**0.192**	**0.415**	NA	**0.532**	†0.415
	ALOPE (LoRMA)	0.365	0.131	0.413	NA	0.512	0.355
Legal	Zero-shot	NA	NA	-0.174	-0.011	-0.429	-0.205
	Few-shot + Guidelines	NA	NA	-0.477	**0.528**	0.304	0.118
	Few-shot (No Guidelines)	NA	NA	-0.459	-0.098	0.032	-0.175
	ALOPE (LoRA)	NA	NA	**0.581**	0.267	**0.445**	†0.431
	ALOPE (LoRMA)	NA	NA	0.515	0.125	0.216	0.265
Tourism	Zero-shot	0.417	0.469	NA	**0.309**	NA	0.398
	Few-shot + Guidelines	-0.459	-0.098	NA	0.032	NA	-0.175
	Few-shot (No Guidelines)	-0.459	-0.098	NA	0.032	NA	-0.175
	ALOPE (LoRA)	0.350	**0.670**	NA	0.220	NA	†0.413
	ALOPE (LoRMA)	**0.465**	0.532	NA	0.227	NA	0.408

Table 10: Comparison of Spearman's correlation (ρ) achieved by LLaMA-3.2-3B across Prompt-only approaches and ALOPE. The bolded values represent the highest Spearman scores obtained for each language pair in each domain. The (†) represents the highest average obtained in each domain. 'NA' indicates that a language pair is unavailable for that specific domain.