Predict and Fix: A Unified Model for Translation Quality Estimation and Post-Editing

Maciej Modrzejewski maciej.modrzejewski@biglanguage.com
BIG Language Solutions

Yash Bhaskar yash.bhaskar@biglanguage.com
BIG Language Solutions

Chinmay Pateria chinmay.pateria@biglanguage.com
BIG Language Solutions

Abstract

We propose a unified architecture for jointly modeling Translation Quality Estimation (QE) and Automatic Post-Editing (APE) within a single lightweight language model. Our approach integrates quality prediction and correction generation in a single decoding process using a decoder-only Qwen2.5 model (0.5B parameters), augmented with a dedicated QE regression head operating on hidden states at a special token position. The model produces structured outputs that include a continuous quality score, an edit decision, and a corrected translation when necessary. We train on datasets of 100K, 1M, and 1.84M manually annotated samples across eight language pairs, enabling analysis of both data scale and distribution. Experimental results show that the proposed model achieves strong QE performance ($r = 0.907$) and high post-editing decision accuracy (88.4%), while reducing over-editing compared to both autoregressive baselines and large commercial LLMs.

1 Introduction

Quality Estimation (QE) provides reference-free assessment of machine translation (MT) output and is widely used in localization workflows to determine whether a translation can be published as-is or should be revised (Zerva et al., 2024). When revision is required, the segment is forwarded to either a human post-editor or an Automatic Post-Editing (APE) system, which in recent deployments is typically a large language model (LLM), often combined with Retrieval-Augmented Generation (Lewis et al., 2020) for terminology and domain conditioning.

In this conventional setup, QE and APE are implemented as two independent models invoked sequentially. The QE model returns a scalar quality score, a thresholding rule converts the score into a routing decision, and a separate APE model produces the corrected output. This decoupling has two consequences. First, the internal representations that the QE model uses to detect translation errors are discarded before post-editing begins, so the APE model must recover comparable information from the same input (Ranasinghe et al., 2020). Second, two forward passes through large models are required for every segment.

We study whether QE and APE can be performed jointly within a single decoder-only model, such that a single forward pass yields (i) a continuous quality score, (ii) PE action and (iii) a corrected translation when PE is needed. The model is built on Qwen2.5-0.5B (Qwen et al., 2025) and is extended with a regression head attached to the hidden state at a designated `<QE>` token position whereas the language modeling head produces the edit decision and post-edited text. The two objectives are optimized together using an uncertainty-weighted multi-task loss (Kendall et al., 2018).

We evaluate this design on manually annotated QE+APE data covering eight English-source language pairs at three training scales (100K, 1M, and

1.84M segments). The contributions of this work are:

- A unified decoder-only architecture that produces a calibrated QE score, an edit decision, and a post-edited translation in a single decoding pass.
- An empirical comparison against commercial LLMs and a decoupled QE+APE pipeline across eight language pairs, including an analysis of over and forced editing behavior.
- A study on the effect of scaling training data and distribution of language pairs on joint QE+APE performance.

2 Related Work

2.1 Reference-free Quality Estimation

Sentence-level QE has been dominated by learned metrics built on multilingual pre-trained encoders. COMET (Rei et al., 2020a) introduced a regression framework over cross-lingual representations, and TransQuest (Ranasinghe et al., 2020) showed that fine-tuned cross-lingual transformers showed gains in representations effectively for QE across language pairs. Subsequent systems such as xCOMET (Guerreiro et al., 2023) and MetricX-23 (Juraska and Others, 2023) extend this line with finer-grained error supervision and report strong correlation with human judgments at the WMT shared tasks (Zerva et al., 2024). In parallel, prompt-based methods including GEMBA (Kocmi et al., 2023; Kocmi and Federmann, 2023) obtain quality assessments directly from large language models. These approaches produce a quality score but do not generate a corrected translation, so any downstream revision still requires a separate model.

2.2 LLM-based Automatic Post-Editing

A second line of work uses large language models for direct post-editing of MT outputs (Zhang et al., 2023; Raunak and Others, 2023; Brown et al., 2020). Such systems are typically prompted in a zero or few shot manner with the source segment and the MT hypothesis and asked to return a revised translation. Variants include error-annotation guided post-editing (Ki and Carpuat, 2024), iterative refinement (Chen and Others, 2023), and self-refinement (Feng and Others, 2024). A repeated observation in this literature is that generative post-editors lack a calibrated decision signal for whether an edit is necessary, which leads to unnecessary modifications of acceptable translations, Huang et al. (2024) further report that LLMs do not reliably self-correct without an external signal.

2.3 Joint and Quality-informed Post-Editing

WMT 2024 Quality-informed Automatic Post-Editing track shared tasks have explicitly targeted the integration of QE and APE also it required systems to combine quality prediction with correction, and quality-aware decoding methods. Most existing systems still couple the two stage process through prompting or pipeline orchestration rather than a unified approach. In contrast, this work performs QE and APE jointly inside a single decoder-only model, with quality estimation produced by a regression head over a designated token. In the same decoding pass the post-edited output is produced , and evaluates the design across eight language pairs and three training scales.

3 Methodology

3.1 Problem Formulation

We formulate joint Quality Estimation and Post-Editing as a single conditional prediction problem. Given a source sentence s and an MT hypothesis m_0, the model jointly predicts a quality score $\hat{q} \in [0, 1]$, a binary edit decision $a \in \{\text{No PE}, \text{PE}\}$, and a post-edited translation f (defined only when $a = \text{PE}$):

$$(\hat{q}, a, f) = \mathcal{F}_\theta(s, m_0). \tag{1}$$

All three outputs are obtained from a single forward pass of the same model, without additional prompting steps or external thresholding (Figure 1).

The QE supervision signal is derived from the normalized edit distance (NED) between the MT output and the linguist correction:

$$\text{NED} = \frac{\text{Lev}(\text{MT}, \text{PE})}{\max(|\text{MT}|, |\text{PE}|)}, \qquad q = 1 - \text{NED}, \tag{2}$$

clipped to $[0, 1]$. Higher q corresponds to fewer required edits.

We use NED as a practical proxy for sentence-level QE because it directly measures the amount of human post-editing required to transform the MT hypothesis into the approved linguist correction. In localization workflows, this edit effort is closely aligned with translation quality: outputs requiring

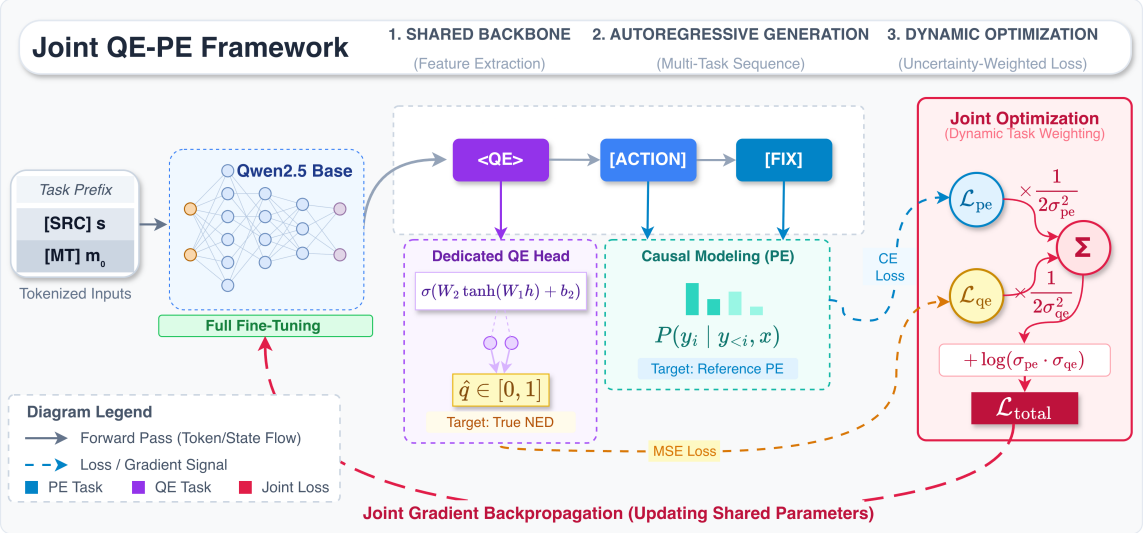

Figure 1: Joint QE+APE architecture. Given a source sentence and an MT hypothesis, the Qwen2.5 backbone autoregressively decodes a single output sequence containing the <QE> token, an edit decision (No PE / PE), and a post-edited translation when applicable. The final-layer hidden state at the <QE> position is read once and passed through a regression head to produce the scalar quality score $\hat{q}$. The language modeling and regression objectives are optimized jointly with uncertainty-weighted losses; no additional model calls or external thresholding are required at inference.

fewer edits are generally more acceptable for publication, whereas outputs requiring substantial edits indicate lower quality. Although NED does not capture all semantic adequacy or fluency issues perfectly, it provides a consistent, reference-based supervision signal that is available for every post-edited segment and is therefore suitable for training a calibrated QE regression model.

3.2 Architecture

The backbone is a decoder-only language model in our experiments we use Qwen2.5-0.5B (Qwen et al., 2025), but the formulation applies to any autoregressive transformer. The vocabulary is extended with a special token <QE>, following the use of dedicated control tokens in prior quality-aware work (Koneru et al., 2025). During training, the input sequence is constructed so that the <QE> token precedes the structured output containing the edit decision and the post-edited translation. The hidden state at the position of this token is used as the input to the QE regression head.

Let $h_{<QE>} \in \mathbb{R}^d$ denote the final-layer hidden state at the special token <QE> position. The QE score is obtained by a two-layer regression head:

$$\hat{q} = \sigma(\mathbf{W}_2 \cdot \text{Dropout}(\tanh(\mathbf{W}_1 h_{<QE>} + b_1)) + b_2), \quad (3)$$

with $\mathbf{W}_1 \in \mathbb{R}^{d \times d}$, $\mathbf{W}_2 \in \mathbb{R}^{1 \times d}$. The sigmoid output constrains $\hat{q}$ to $[0, 1]$ to match the NED-derived target. The language modeling head of the main pre-trained model is unchanged and produces the structured output token sequence containing the action label and, when applicable, the corrected translation.

Because the regression head consists of only two linear layers operating on an already-computed hidden state, it adds negligible computational overhead during inference.

3.3 Training Objective

The model is trained with two losses computed on the same forward pass: a regression loss $\mathcal{L}_{qe}$ (mean squared error between $\hat{q}$ and the NED-derived target) and the standard cross-entropy language modeling loss $\mathcal{L}_{text}$ over the structured output tokens. The two losses are combined using homoscedastic uncertainty weighting (Kendall et al., 2018), with learnable log-variance parameters s_{qe} and s_{text}:

$$\mathcal{L}_{joint} = e^{-s_{qe}} \mathcal{L}_{qe} + e^{-s_{text}} \mathcal{L}_{text} + s_{qe} + s_{text}. \quad (4)$$

The exponential representation keeps the task weights positive, and the additive s_{qe} and s_{text} terms prevent the loss from being minimized by increasing the log-variances without bound. Both s_{qe} and s_{text} are updated jointly with the model parameters during training.

4 Experimental Setup

4.1 Data

We construct three training datasets of approximately 100K, 1M, and 1.84M segments covering eight English-source language pairs: French (Canadian), Japanese, Spanish, Korean, Simplified Chinese, German, Polish, and European French. For every segment, the MT hypothesis is produced by an in-house translation system fine-tuned for the target domain, and the reference is a post-edited correction approved by professional linguists (gold references). The 100K and 1M datasets follow the natural distribution of the source corpus, whereas the 1.84M dataset is rebalanced to allocate approximately 250K segments per language pair, with the exception of English to European French (93,364 segments), for which fewer instances were available. The per-pair distribution for each scale is reported in Table 1.

For evaluation, we hold out a test set of 9,048 segments sampled from the same distribution as the training data. Table 2 reports BLEU and COMET scores between the in-house MT output and the gold references on this test set. Scores are high across pairs (BLEU above 60 on most pairs, COMET close to or above 1.0 on French, Spanish, and Korean), which limits the room for surface-metric improvement through post-editing on the higher-resource pairs.

Each segment is annotated with a binary decision (PE or No PE) and a corrected translation when applicable. The continuous regression target is $q = 1 - \text{NED}$, where NED is the length-normalized Levenshtein distance between the MT hypothesis and the gold reference (Section 3).

4.2 Baselines

We compare the proposed joint model against three baselines. The first is an autoregressive variant of our own model, and the remaining two are zero-shot commercial LLMs. All three are evaluated on the same 9,048 segment test set.

Autoregressive QE+APE. The same Qwen2.5-0.5B model is fine-tuned without the regression head. The QE score is generated as part of the output sequence rather than predicted from the `<QE>` hidden state. This baseline isolates the contribution of the regression head when training data is the same. We train this baseline on the 100K split.

GPT-5.4 and **Gemini 3.1 Pro** are evaluated in a zero-shot setting. Since we cannot access the hidden state of these models for the regression head, we prompt these models to output a numeric score with an action and a corrected translation if applicable. The full prompt used here is shown below.

Zero-Shot LLM Baseline Prompt

SYSTEM: You are a translation quality judge from {src_lang} into {tgt_lang}.
Given only a source sentence (SOURCE) and a machine translation (MT), decide if any post-editing (PE) is needed.
Assign a QE score between 0 and 100 where:
100 = perfect translation, no errors
0 = completely incorrect translation.
If no PE is needed, output exactly:

```
QE: <score between 0 and 100>
ACTION: No Post-Edition
```

If Post-Edition is needed, output exactly:

```
QE: <score between 0 and 100>
ACTION: Post-Edition
FIX: <the corrected translation>
```

Always return only these fields in this order.

USER:

```
[LANGPAIR] {src_lang}-{tgt_lang}
[SOURCE] {src_text}
[MT] {mt_text}
```

Language Pair	BLEU	COMET	#Samples
eng→fr_ca	85.62	1.062	3706
eng→jap	69.21	0.808	1067
eng→spa	82.56	1.046	1001
eng→kor	77.18	0.991	942
eng→chi_s	66.13	0.757	906
eng→ger	69.17	0.793	732
eng→pol	62.85	0.978	576
eng→fre	61.15	0.869	118

Table 2: Baseline MT quality on the test set across language pairs, measured in BLEU and COMET, with the number of test samples per pair.

Language Pair	100K %	100K Segments	1M %	1M Segments	1.84M %	1.84M Segments
eng→fr_ca	36.78%	36,780	36.69%	366,900	13.56%	250,000
eng→jap	10.99%	10,990	11.03%	110,300	13.56%	250,000
eng→spa	10.33%	10,330	10.31%	103,100	13.56%	250,000
eng→kor	9.82%	9,820	9.86%	98,600	13.56%	250,000
eng→chi_s	8.84%	8,840	8.80%	88,000	13.56%	250,000
eng→ger	7.73%	7,730	7.69%	76,900	13.56%	250,000
eng→pol	6.13%	6,130	6.08%	60,800	13.56%	250,000
eng→fre	1.22%	1,220	1.24%	12,400	5.06%	93,364
Total		100,000		1,000,000		1,843,364

Table 1: Distribution of language pairs across the 100K, 1M, and 1.84M datasets with corresponding segment counts.

Metric	GPT-5.4	Gemini 3.1 Pro
Global QE Performance		
Pearson r	0.2363	0.3096
Spearman ρ	0.3706	0.3314
MAE	0.1656	0.1515
RMSE	0.2775	0.2599
Global PE Decision		
Accuracy	0.6584	0.6132
Precision	0.7427	0.7972
Recall	0.6805	0.4976
F1 Score	0.7102	0.6127

Table 3: Global QE and PE decision performance for commercial LLM baselines.

Unified QE+APE Prompt (Ours)

SYSTEM: You are a translation quality judge from {src_lang} into {tgt_lang}.
Given only a source sentence (SOURCE) and a machine translation (MT), decide whether post-editing (PE) is needed.
If no PE is needed, output exactly:
`QE: <QE>`
`ACTION: No Post-Edition`
If PE is needed, output exactly:
`QE: <QE>`
`ACTION: Post-Edition`
`FIX: <the corrected translation>`
Always return only these fields in this order.

USER:
`[LANGPAIR] {src_lang}-{tgt_lang}`
`[SOURCE] {src_text}`
`[MT] {mt_text}`

4.3 Proposed Architecture

All fine-tuning experiments for the proposed method uses Qwen2.5-0.5B (Qwen et al., 2025) model, with a regression head described in Section 3 attached to the hidden state at the<QE> token position. The model is trained with the structured output prompt shown below, which fixes the <QE> token in the output.

4.4 Evaluation Metrics

We report three groups of metrics.

Quality Estimation. We compute Pearson's r and Spearman's ρ between predicted QE scores and the gold scores. We also report Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) between them.

PE Decision Classification. For the PE or No PE prediction, we report Accuracy, Precision, Recall, and F1. We additionally report the accuracy on the No PE subset, which measures how often the model reports no PE required for an acceptable translation.

PE Quality. For segments on which the model decides to PE, we measure the ΔBLEU and ΔCOMET (`Unbabel/wmt20-comet-da`) (Rei et al., 2020b) for translation quality relative to the original MT output.

5 Results and Analysis

We organize the analysis around four research questions: (i) Does the QE regression head improve quality prediction over an autoregressive QE baseline trained on the same data? (ii) How do QE and APE performance scale with training set size and language-pair balancing? (iii) How does the proposed model compare with zero-shot commercial LLMs? (iv) How does post-editing quality vary

Lang Pair	GPT-5.4				Gemini 3.1 Pro			
	PE Acc	QE r	ΔBLEU	ΔCOMET	PE Acc	QE r	ΔBLEU	ΔCOMET
eng→fr_ca	0.656	0.293	-15.25	-0.016	0.616	0.281	-20.27	-0.059
eng→jap	0.658	0.223	-1.46	-0.006	0.577	0.266	+6.76	-0.024
eng→spa	0.618	0.124	-17.69	-0.021	0.613	0.338	-23.10	-0.059
eng→kor	0.679	0.337	-11.85	-0.001	0.623	0.423	-7.42	-0.021
eng→chi_s	0.648	0.251	+6.84	+0.004	0.622	0.315	+5.56	-0.011
eng→ger	0.677	0.310	-10.74	-0.013	0.659	0.336	-15.90	-0.047
eng→pol	0.693	0.253	-11.05	-0.003	0.610	0.302	-10.80	-0.042
eng→fre	0.695	0.347	-4.10	-0.019	0.441	0.099	-21.50	-0.190

Table 4: Per-language zero-shot post-editing performance for GPT-5.4 and Gemini 3.1 Pro. ΔBLEU and ΔCOMET indicate the change relative to the original MT output.

Model	Pearson	Spearman	MAE	RMSE	Acc (All)	Acc (No PE)	Acc (PE)
Baseline AR (100K)	0.094	0.076	0.357	0.528	0.798	0.689	0.866
QE+APE (100K)	0.809	0.762	0.078	0.152	0.822	0.751	0.866
QE+APE (1M)	0.863	0.837	0.059	0.125	0.877	0.814	0.914
QE+APE (1.84M)	0.907	0.876	0.054	0.112	0.884	0.777	0.939

Table 5: QE regression performance and APE decision accuracy. Accuracy (All) is the overall binary decision of whether post-editing is needed. Acc (No PE) measures correct identification of sentences that do not require post-editing, while Acc (PE) measures correct identification of sentences that do require post-editing.

across language pairs?

5.1 Effect of the QE Regression Head

Table 5 compares the autoregressive baseline trained on 100K segments with the proposed QE+APE model trained on the same amount of data. Replacing in-sequence QE score generation with a regression head over the <QE> hidden state substantially improves QE performance. Pearson's r increases from 0.094 to 0.809, while RMSE decreases from 0.528 to 0.152.

The regression head also improves PE decision quality. At the same 100K scale, overall PE decision accuracy increases from 79.8% for the autoregressive baseline to 82.2% for the proposed model. The subset accuracies show that the autoregressive baseline achieves 68.9% accuracy on No-PE examples and 86.6% on PE examples, indicating a tendency to over-edit acceptable translations. The proposed model improves No-PE recognition to 75.1% while maintaining the same PE accuracy of 86.6%, suggesting that explicit QE supervision helps the model make more selective post-editing decisions.

5.2 Effect of Training Scale and Language Balance

Increasing the training data improves both QE prediction and PE decision accuracy. As shown in Table 5, Pearson's r increases from 0.809 with 100K samples to 0.863 with 1M samples and 0.907 with 1.84M samples. Similarly, overall PE decision accuracy increases from 82.2% to 87.7% and then to 88.4%. MAE and RMSE also decrease consistently, showing that larger training sets produce better calibrated QE scores.

The per-language post-editing results in Table 6 show a similar trend. On Japanese, Korean, German, Polish, and Simplified Chinese, the larger models generally produce stronger ΔBLEU and ΔCOMET results. In several cases, scaling turns negative or weak post-editing outcomes into positive gains. For example, German improves from -15.37 ΔBLEU with the autoregressive 100K baseline to $+6.17$ with the 1.84M QE+APE model, while Polish improves from -12.57 to $+0.69$.

However, the effect of scale is also influenced

Lang Pair	Baseline AR (100K)		QE+APE (100K)		QE+APE (1M)		QE+APE (1.84M)	
	ΔBLEU	ΔCOMET	ΔBLEU	ΔCOMET	ΔBLEU	ΔCOMET	ΔBLEU	ΔCOMET
eng→fr_ca	+5.10	+0.03	-0.55	-0.01	+8.09	+0.08	+1.49	-0.02
eng→jap	+8.86	+0.24	+5.03	+0.14	+7.65	+0.20	+7.32	+0.17
eng→spa	-20.04	-0.18	-19.57	-0.14	-16.50	-0.15	-15.64	-0.18
eng→kor	+1.23	-0.03	+5.49	+0.06	+15.19	+0.08	+16.72	+0.09
eng→chi_s	+3.71	+0.07	+4.90	+0.12	+5.13	+0.13	+6.68	+0.07
eng→ger	-15.37	-0.15	-9.85	-0.17	-2.30	-0.08	+6.17	+0.02
eng→pol	-12.57	-0.34	-15.27	-0.23	-2.67	-0.09	+0.69	+0.02
eng→fre	-19.62	-0.47	-18.47	-0.32	-10.39	-0.19	-12.11	-0.18

Table 6: Average change in BLEU and COMET after post-editing relative to the original MT output for each language pair and model configuration. Positive values indicate improvements over the original MT, while negative values indicate degradation. Green highlights the best-performing configuration for a given language pair and metric, whereas red marks the worst-performing (most negative) result.

Lang Pair	GPT-5.4		Gemini 3.1 Pro		QE+APE (1M) [Ours]		QE+APE (1.84M) [Ours]	
	ΔBLEU	ΔCOMET	ΔBLEU	ΔCOMET	ΔBLEU	ΔCOMET	ΔBLEU	ΔCOMET
eng→fr_ca	-15.25	-0.016	-20.27	-0.059	**+8.09**	**+0.080**	+1.49	-0.020
eng→jap	-1.46	-0.006	+6.76	-0.024	**+7.65**	**+0.200**	+7.32	+0.170
eng→spa	-17.69	**-0.021**	-23.10	-0.059	-16.50	-0.150	**-15.64**	-0.180
eng→kor	-11.85	-0.001	-7.42	-0.021	+15.19	+0.080	**+16.72**	**+0.090**
eng→chi_s	**+6.84**	+0.004	+5.56	-0.011	+5.13	**+0.130**	+6.68	+0.070
eng→ger	-10.74	-0.013	-15.90	-0.047	-2.30	-0.080	**+6.17**	**+0.020**
eng→pol	-11.05	-0.003	-10.80	-0.042	-2.67	-0.090	**+0.69**	**+0.020**
eng→fre	**-4.10**	**-0.019**	-21.50	-0.190	-10.39	-0.190	-12.11	-0.180

Table 7: Language-level comparison of post-editing impact (ΔBLEU and ΔCOMET). Best results per metric are in **bold**.

by language-pair distribution. The 1.84M dataset is more balanced across language pairs, which improves performance for several lower-resource pairs. At the same time, Canadian French receives a smaller relative share in the 1.84M setting than in the naturally distributed 1M setting, and the 1M model performs better for this pair. This suggests that scaling helps most when it increases both total data volume and sufficient per-language coverage.

5.3 QE and APE Prediction Quality/Accuracy

Table 8 compares our Qwen-based QE+APE models trained on 1M and 1.84M samples against zero-shot GPT-5.4 and Gemini 3.1 Pro. The commercial LLMs obtain PE decision accuracies of 65.8% and 61.3%, respectively, with relatively low QE correlations ($r = 0.236$ and $r = 0.310$). In contrast, the Qwen QE+APE models are substantially stronger: the 1M model reaches 87.7% PE decision accuracy and $r = 0.863$, while the 1.84M model further improves to 88.4% accuracy and $r = 0.907$.

These results show that the proposed Qwen models provide much more reliable QE and APE predictions than the baseline LLMs. The improvement suggests that task-specific joint training with an explicit QE regression head is more effective for calibrated translation-quality prediction than zero-shot prompting of larger general-purpose LLMs.

5.4 Per-Language Post-Editing Quality

Tables 6 and 7 report ΔBLEU and ΔCOMET on the segments where each model selects PE decision.

Within our model family. Across the four scales (Baseline AR 100K, QE+APE at 100K, 1M, and

1.84M), ΔBLEU improves with scale on six of the eight pairs. On simplified chinese, japanese, korean, german, and polish, the model trained on 1.84M segments is the best in its column and turns the negative ΔBLEU values of the smaller scales into positive ones (ger: $-15.37 \rightarrow +6.17$; pol: $-12.57 \rightarrow +0.69$; kor: $+1.23 \rightarrow +16.72$). On spa and fre the trend is the same direction (less negative with scale) but ΔBLEU remains negative even at 1.84M (-15.64 and -12.11). For fr_ca the 1M model is best ($+8.09$ BLEU); the 1.84M model is lower ($+1.49$), consistent with the reduced fr_ca share in the rebalanced training set.

Against commercial LLMs. GPT-5.4 and Gemini 3.1 Pro produce negative ΔBLEU on six of the eight pairs, with the largest drops on fr_ca, fre, and spa (Gemini: -20.27, -21.50, -23.10). The 1.84M model is positive or close to zero on six of eight pairs and is the best column on five (ger, kor, pol, jap, chi_s tied). The persistent negative ΔBLEU on spa and fre is shared with both commercial models, indicating that the difficulty is not specific to model capacity. We discuss this further in Section 5.5.

5.5 Over-Editing

Two language pairs, eng→spa and eng→fre, show negative ΔBLEU and ΔCOMET across all model configurations, including the 1.84M QE+APE model and both commercial LLMs. For these pairs, the original in-house MT output is already strong, with BLEU scores of 82.56 and 61.15 and COMET scores of 1.046 and 0.869, respectively, as shown in Table 2. When the model edits translations that are already close to the reference, even small surface-level changes can reduce BLEU or COMET.

The No-PE subset accuracy in Table 5 further supports this interpretation. The 1.84M model achieves 77.7% accuracy on No-PE examples, meaning that it still edits a non-trivial fraction of translations that should have been left unchanged. This behavior reflects the central challenge of QE-informed post-editing: the model must not only generate good corrections, but also learn when no correction is needed.

The same pattern appears in the commercial LLM baselines, which often produce larger negative ΔBLEU values on high-quality language pairs. Compared with these baselines, our QE+APE models generally reduce the severity of over-editing, but further calibration of the PE decision threshold remains an important direction for future work.

Model	QE r	QE MAE	PE Acc
GPT-5.4	0.236	0.165	0.658
Gemini 3.1 Pro	0.310	0.151	0.613
QE+APE (1M) [Ours]	0.863	0.059	0.877
QE+APE (1.84M) [Ours]	**0.907**	**0.054**	**0.884**

Table 8: Global comparison between commercial LLMs and our proposed joint QE+APE models (1M and 1.84M samples).

6 Limitations

While the proposed unified QE+APE architecture demonstrates strong performance, we acknowledge certain limitations in our current experimental setup.

First, our evaluation compares a model fine-tuned on in-domain data against commercial LLMs (GPT-5.4 and Gemini 3.1 Pro) operating in a zero-shot setting. Because the available general purpose models did not have access to the specific domain distribution of our training data, the comparison is inherently asymmetric.

Second, due to resource constraints, our autoregressive baseline was trained only at the 100K scale. While this serves as a proof ofconcept that explicitly modeling the regression head improves performance over standard autoregressive decoding.

Third, the training targets and test references are derived directly from corrections made by professional linguists. This ensures our Normalized Edit Distance targets accurately represent real world human PE effort. However, evaluation of our generated outputs relies on standard reference based automated metrics (BLEU and COMET).

Finally, regarding reproducibility, the datasets used in this study consist of proprietary in-house data and thus cannot be publicly released. However, our architectural setup and prompt designs are fully detailed to enable replication on public datasets.

7 Future Work

We identify two directions for future work First, re-evaluate available commercial LLMs with few-shot prompting would narrows the data gap without changing their architecture, so any remaining performance difference can be attributed to the unified

QE+APE approach rather than to domain familiarity alone. Second, Extending the head to predict structured like MQM style annotations, error spans, categories, and severities jointly with the correction would scale from a scalar acceptability judgment toward interpretable, error-localized feedback that is more directly useful to linguists.

8 Conclusion

We have shown that jointly fine-tuning a small, specialized language model is an effective approach for translation QE and APE. By integrating quality estimation and post-editing within a single architecture, the model can use a shared representation of translation quality to guide both score prediction and correction generation. This enables more controlled post-editing than a pipeline or zero-shot prompting setup.

Our results show that task-specific joint training is more important than model scale for this setting. The Qwen-based QE+APE models substantially outperform the commercial LLM baselines in prediction accuracy: the 1M and 1.84M models achieve PE decision accuracies of 87.7% and 88.4%, respectively, compared with 65.8% for GPT-5.4 and 61.3% for Gemini 3.1 Pro. The 1.84M model also reaches a much stronger QE correlation, with Pearson's $r = 0.907$, showing that the proposed regression-head design produces more reliable quality estimates than zero-shot general-purpose LLMs.

Overall, the proposed QE+APE framework reduces the over-editing commonly observed in non-specialized models and provides a more reliable foundation for automatic post-editing in localization workflows. Future work should focus on improving calibration and reducing unnecessary edits, especially for language pairs where the original MT output is already strong.

References

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Chen, P. and Others (2023). Iterative translation refinement with large language models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation (EAMT)*.

Feng, Y. and Others (2024). TEaR: Improving LLM-based machine translation with self-refinement. In *Findings of the Association for Computational Linguistics: NAACL 2024*.

Guerreiro, N. M., Rei, R., and Others (2023). xCOMET: Transparent machine translation evaluation through fine-grained error detection. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*.

Huang, J., Chen, X., Mishra, S., Zheng, H. S., Yu, A. W., Song, X., and Chi, E. H. (2024). Large language models cannot self-correct reasoning yet. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Juraska, J. and Others (2023). MetricX-23: The google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*.

Kendall, A., Gal, Y., and Cipolla, R. (2018). Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491.

Ki, C. and Carpuat, M. (2024). Guiding large language models to post-edit machine translation with error annotations. In *Findings of the Association for Computational Linguistics: NAACL 2024*.

Kocmi, T., Bawden, R., Bojar, O., Martins, A. F. T., et al. (2023). Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation (EAMT)*.

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

Koneru, S., Huck, M., Exel, M., and Niehues, J. (2025). Quality-aware decoding: Unifying quality estimation and decoding. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 33–46.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Qwen, :, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., and Qiu, Z. (2025). Qwen2.5 technical report.

Ranasinghe, T., Orasan, C., and Mitkov, R. (2020). Transquest: Translation quality estimation with cross-lingual transformers. In *Proceedings of the 28th international conference on computational linguistics*, pages 5070–5081.

Raunak, V. and Others (2023). Leveraging large language models for automated machine translation post-editing. *arXiv preprint arXiv:2305.xxxx*.

Rei, R., Stewart, C., Farinha, A. C., and Lavie, A. (2020a). Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)*, pages 2685–2702.

Rei, R., Stewart, C., Farinha, C., and Lavie, A. (2020b). Unbabel's participation in the wmt20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Zerva, C., Blain, F., De Souza, J. G., Kanojia, D., Deoghare, S., Guerreiro, N. M., Attanasio, G., Rei, R., Orasan, C., Negri, M., et al. (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*, pages 82–109.

Zhang, B., Haddow, B., and Birch, A. (2023). Prompting large language model for machine translation: A case study. In *International conference on machine learning*, pages 41092–41110. PMLR.