CompactQE: Interpretable Translation Quality Estimation via Small Open-Weight LLMs

Kamil Guttmann[1,2], **Zofia Fraś**[1], **Artur Nowakowski**[1,2], **Krzysztof Jassem**[1,2]
[1] Laniqo, Poznań, Poland
[2] Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poznań, Poland
{name}.{surname}@laniqo.com

Abstract

Current state-of-the-art Quality Estimation (QE) in machine translation relies on massive, proprietary LLMs, raising data privacy concerns. We demonstrate that smaller, open-source LLMs (<30B parameters) are a viable, cost-effective and privacy-preserving alternative. Using a single-pass prompting strategy, our models simultaneously generate quality scores, MQM error annotations, suggested error corrections, and full post-editions. Our analysis shows these models achieve highly competitive system-level correlations with human judgments that outperform traditional neural metrics, fine-tuned models, and human inter-annotator agreement, effectively approximating the capabilities of much larger proprietary LLMs.

1 Introduction

Quality Estimation (QE) remains a critical component in professional machine translation workflows, enabling the assessment of translation reliability without access to reference translations. While neural regression-based metrics such as COMET (Rei et al., 2020) have achieved high correlations with human judgments, they typically produce a single scalar score. This aggregation provides little insight into specific translation errors or their severity, limiting the utility of such metrics for downstream tasks that require detailed feedback and interpretability. Consequently, there is a growing demand for evaluation methods that can offer fine-grained error diagnosis alongside quality scores.

The emergence of Large Language Models (LLMs) has introduced a new paradigm for evaluation, often referred to as "LLM-as-a-judge". Approaches such as GEMBA (Kocmi and Federmann, 2023b) and AutoMQM (Fernandes et al., 2023) leverage the generative capabilities of LLMs to mimic human evaluation frameworks like Multidimensional Quality Metrics (MQM) (Lommel et al., 2013) or Error Span Annotation (ESA) (Kocmi et al., 2024). By employing zero-shot or few-shot prompting strategies, these methods can identify and categorize error spans, offering a level of interpretability that was previously difficult to attain with standard neural metrics.

However, the state-of-the-art performance of these generative metrics currently relies heavily on massive, proprietary models such as GPT-5 or Gemini. This dependency presents significant challenges for professional deployment, particularly concerning data privacy. Sending sensitive content to external APIs for evaluation is often prohibitive for enterprises with strict data security requirements. Furthermore, the lack of transparency regarding the training data and architecture of closed models raises concerns about reproducibility and the stability of evaluation standards over time.

We hypothesize that high-quality, interpretable QE can be achieved without reliance on very large proprietary models. In this work, we demonstrate the feasibility of developing an effective MQM-style metric using significantly smaller, open-source multilingual models. By shifting to open weights, we address the critical need for privacy-preserving evaluation pipelines that can be deployed locally within secure environments. We

show that with appropriate prompting and output processing, these smaller models can offer more efficient alternatives to their larger, closed-source counterparts.

To achieve competitive performance with these resource-constrained architectures, we introduce a streamlined methodology that integrates quality estimation and post-editing. Building upon the structured prompting strategies of GEMBA-MQM V2 (Junczys-Dowmunt, 2025), our approach performs translation post-editing and error detection in a single inference step. This contrasts with recent multi-stage approaches like MQM-APE (Lu et al., 2025), which require multiple model calls to verify errors. We subsequently apply heuristic error filtering to refine the output, reducing computational overhead while maintaining the granularity required for professional quality checks.

2 Related Work

The COMET-Kiwi (Rei et al., 2022) family of metrics and recent models like MetricX-25-QE (Juraska et al., 2025) represent state-of-the-art neural regression baselines explicitly fine-tuned for reference-free translation quality estimation. These models leverage pretrained multilingual encoders to predict a single quality score. While the scalar score can be a good indicator of the translation's quality, it is not easily interpretable.

xCOMET (Guerreiro et al., 2024) bridges the gap between regression-based scoring and fine-grained error detection by integrating both tasks into a single learned metric. It predicts sentence-level scores while simultaneously identifying error spans and assigning them severity labels (Minor, Major, Critical), achieving high correlation with human scores. While it is a big leap from scalar metrics it still lacks the fine-grained error categorization found in human-based error annotations such as MQM.

Generative, LLM-based approaches like Error Analysis Prompting (EAPrompt) (Lu et al., 2024) attempt to emulate the human MQM framework by combining Chain-of-Thought (CoT) reasoning with structured error analysis. EAPrompt explicitly instructs the model to identify major and minor errors within its reasoning steps before calculating a final score. While the internal CoT process identifies specific errors, the primary output focus remains the aggregated score or error count, rather than returning a structured list of error spans to the user. Furthermore, EAPrompt employs multi-turn prompting, separating error identification from error counting, which increases the computational cost by requiring multiple model calls per evaluated segment.

AutoMQM represents one of the initial efforts to fully automate the annotation of translations with MQM-style errors using Large Language Models. By leveraging large models like PaLM-2, AutoMQM demonstrated that LLMs could effectively identify and categorize errors when provided with few-shot examples sampled from human annotations. This approach shifted the focus from assigning a scalar score to generating structured error annotations.

The GEMBA family of metrics established the state-of-the-art for LLM-based evaluation using zero-shot and few-shot prompting with GPT models. GEMBA-MQM (Kocmi and Federmann, 2023a) specifically prompts GPT-4 to output structured error annotations, mapping severities to a final quality score. The most recent iteration, GEMBA-MQM V2, addresses the stochastic nature of LLM outputs by aggregating scores across ten distinct inference runs. While this aggregation strategy significantly improves reliability and correlation with human judgments, it increases computational demands, requiring large proprietary models and substantially higher inference time and cost per segment.

MQM-APE builds upon the generative evaluation paradigm by integrating an Automatic Post-Editing (APE) step to verify the impact of detected errors. It filters out non-impactful errors by checking if fixing them actually improves the translation quality according to a pairwise verifier. Notably, MQM-APE demonstrated the viability of using open-source models like Llama-3 (Grattafiori et al., 2024) and Mixtral (Jiang et al., 2024) for this task. However, its architecture relies on a sequential pipeline of three distinct modules: evaluator, post-editor, and verifier. Since all of these modules are operated by the LLM, it results in high inference costs.

Our work builds directly upon the structured prompting strategies of GEMBA-MQM V2 but adapts them for resource-constrained environments. We extend the framework by incorporating simultaneous post-editing and ESA scoring within a single inference step, alongside more detailed error descriptions. Unlike GEMBA-MQM

V2, which utilizes full document context, we omit document-level information to accommodate the context window limitations of smaller open-source models deployed on consumer-grade GPUs (up to 32GB of VRAM). This trade-off allows us to maintain high interpretability and accuracy while ensuring data privacy and reducing computational overhead.

3 Methodology

3.1 Data

For all experiments we utilize the official data and human annotations provided by the organizers of the WMT25 Metrics Shared Task. The dataset comprises paragraph-level translations generated by various systems submitted to the WMT25 General MT Task. Each of the system outputs was originally evaluated by the task organizers or participants using multiple automated metrics, including COMET22, COMETKiwi22, MetricX-25-QE, GEMBA-v2 and Gemini2.5-Pro, which serve as the baselines for our proposed approach. Subsequently, the organizers selected a subset of the best scoring systems for human evaluation.

Human evaluation followed the ESA protocol, which has been established as the standard evaluation framework at WMT for the past two iterations. According to the ESA guidelines, human evaluators assess translation quality through a two-step process. First, the annotators identify and highlight specific error spans within the translated text, categorizing them by severity. Second, after completing the span annotation, the evaluators assign an overall quality score ranging from 0 to 100 to the translation. Each segment was evaluated by two annotators. We take the average of the two scores and treat it as the golden standard.

From the complete WMT25 dataset, we select three language pairs to evaluate our method across diverse linguistic scenarios: Czech-German (to represent a non-English-centric translation direction), English-Italian (representing a high-resource language pair), and English-Ukrainian (representing a low-resource target language utilizing a non-Latin script). To ensure a fair and consistent comparison between human judgments and automated metrics, we restrict our experiments strictly to the segments that possess complete human ESA annotations and scores. The resulting number of systems and segments per system is reported in Table 1.

Czech → German	
Number of systems	20
Segments per system	231
English → Italian	
Number of systems	18
Segments per system	215
English → Ukrainian	
Number of systems	18
Segments per system	199

Table 1: Details on the number of systems and segments per language pair. Each segment consisted of multiple sentences, making the evaluation paragraph-level.

3.2 Prompts

To query the models for QE, we base our initial prompt structure on GEMBA-MQM V2. This method was among the best-performing metrics in the WMT25 Metrics Task and provides an effective baseline framework for extracting structured error span annotations, with both input and output formatted as JSON.

In its original formulation, the GEMBA-MQM V2 prompt includes the context of the entire document to score individual segments. While proprietary models generally handle large inputs without issues, many small open-source models possess limited context windows. Furthermore, expanding the active context length significantly increases VRAM consumption. This is often impractical for on-premise deployments reliant on consumer-grade hardware due to the associated infrastructure costs and reduced inference speeds. Consequently, we omit the document-level context in our approach, evaluating each paragraph in isolation to fit these hardware constraints.

To calibrate the model's sensitivity and encourage correct JSON output formatting, we provide in-context learning examples within the prompt. By supplying few-shot examples of segments containing high, medium, and zero error counts, we aim to mitigate common issues such as over-annotation (hallucinating non-existent errors) and under-annotation (missing clear mistakes). The examples were specifically selected from translations across multiple language pairs (English-German, Polish-Spanish, French-Italian) to prevent the model from focusing on pair-specific features. This results in a language-pair-independent prompt, which can be applied to any translation pair without the need to modify or replace exam-

ples for subsequent experiments.

We modify the prompt to explicitly request the extraction of both source and target error phrases. Capturing the exact target phrase allows for error span annotation within the translated text. Simultaneously, extracting the corresponding source phrase is necessary for identifying omission errors, where a specific phrase present in the source is entirely absent from the translation.

Alongside the error span annotations, the prompt instructs the model to generate a suggested correction for each identified mistake. Providing targeted fixes has practical utility in professional translation workflows. It can accelerate the human post-editing process by allowing translators to review and accept specific suggestions directly in their computer-assisted translation (CAT) tools.

Finally, we ask the model to output a full post-edited version of the target segment within the same prompt. By consolidating QE, error span annotation and automatic post-edition into a single prompt, we aim to significantly reduce latency and computational overhead.

We provide the system and user prompt templates used for all our experiments in Appendix A, Listings 1, and 2.

3.3 Models

We evaluate our approach using three open-source models: `gemma-3-27b-it`[1] (Team et al., 2025), `EuroLLM-9B-Instruct`[2] (Martins et al., 2025), and `Qwen3-VL-30B-A3B-Instruct`[3] (Bai et al., 2025). We selected these specific models because they fall under the 30-billion parameter threshold, enabling on-premise deployment on smaller, consumer-grade GPUs. Furthermore, they are available under non-restrictive open-source licenses and possess strong multilingual capabilities. Although `Qwen3-VL-30B-A3B-Instruct` is inherently a Vision-Language (VL) Model, we utilize only its text-processing modalities for this task, as it is reported to achieve superior results on text-only benchmarks compared to its non-VL counterpart (Bai et al., 2025).

To benchmark our results against state-of-the-art proprietary systems, we additionally run quality estimation using `Gemini-3-Flash` (Google DeepMind, 2025) as a backbone model. All models are prompted using an identical prompt structure.

For the open-source models, we apply greedy decoding by setting the temperature to zero (`T = 0`). Conversely, we run `Gemini-3-Flash` with `T = 1`, as decreasing the temperature is not recommended in its documentation. To ensure a fair comparison, we minimize the impact of Gemini's internal reasoning capabilities by setting its thinking budget to zero. While the Gemini 3 family does not support a complete 'thinking-off' mode, this configuration triggers a 'minimal' thinking level, effectively preventing the model from complex reasoning.

3.4 Error Filtering

The raw output from generative models often contains inconsistencies, necessitating a structured filtering pipeline. To ensure robust parsing of the generated responses, we utilize the `json-repair` library[4] (Baccianella, 2025) to automatically correct any malformed JSON structures before further processing.

Our first filtering step addresses error span hallucinations, where the model identifies a target phrase that does not actually exist in the generated translation. We filter out these errors based on exact string matching after lowercase normalization.

Furthermore, we observe that models frequently duplicate the same error across multiple severity categories. In these instances, we deduplicate the annotations by retaining only the instance with the highest severity. This heuristic ensures that critical or major translation flaws are accurately represented and not overshadowed or mistakenly downgraded in favor of redundant minor complaints. The filtering results for each model are presented in Table 2.

Additionally, we explored a secondary filtering mechanism based on the generated post-editions. This approach used fuzzy matching to compare the model's suggested correction with its post-edited segment. The underlying logic assumes that if the model flags an error and proposes a fix, but fails to incorporate that fix into its final post-edition, the flagged error is likely a false positive. However,

[1] https://huggingface.co/google/gemma-3-27b-it

[2] https://huggingface.co/utter-project/EuroLLM-9B-Instruct

[3] https://huggingface.co/Qwen/Qwen3-VL-30B-A3B-Instruct

[4] https://github.com/mangiucugna/json_repair

	Error count		Rejection rate
	Before filtering	After filtering	
Czech → German			
EuroLLM-9B-Instruct	20,278	13,591	33.0%
Qwen3-VL-30B-A3B-Instruct	44,099	21,871	50.4%
Gemma-3-27b-it	12,467	10,234	17.9%
Gemini-3-Flash	16,247	15,134	6.9%
English → Italian			
EuroLLM-9B-Instruct	11,387	7,953	30.2%
Qwen3-VL-30B-A3B-Instruct	30,119	16,530	45.1%
Gemma-3-27b-it	8,407	7,351	12.6%
Gemini-3-Flash	13,519	13,032	3.6%
English → Ukrainian			
EuroLLM-9B-Instruct	12,212	8,180	33.0%
Qwen3-VL-30B-A3B-Instruct	24,277	15,357	36.7%
Gemma-3-27b-it	9,509	7,996	15.9%
Gemini-3-Flash	15,324	14,609	4.7%

Table 2: Total number of errors detected by four QE models across three language pairs, before and after filtering, with rejection rates.

our preliminary tests yielded a low level of improvement when applying this filter. We conclude that this heuristic relies heavily on the model possessing exceptionally strong and consistent post-editing capabilities.

This observation aligns closely with the findings from the WMT25 Metrics Task 3, which focused on post-editing sentences with pre-annotated error spans. Lavie et al. (2025) noted that while LLMs demonstrate an ability to improve traditional Neural Machine Translation (NMT) outputs, they struggle to meaningfully improve outputs already generated by LLMs. Consequently, we do not include this filtering step, as the current post-editing limitations of smaller open-source models can introduce noise into the filtering process rather than refine it.

4 Evaluation

Following the evaluation framework implemented in the `mt-metrics-eval`[5] toolkit, we compute correlations between our QE scores, selected baseline metrics, and human ESA ratings. For the system-level evaluation, we report Soft Pairwise Accuracy (SPA) (Thompson et al., 2024). At the segment level, we evaluate the metrics using the "group-by-item" segment-level accuracy with tie calibration (Deutsch et al., 2023).

[5] https://github.com/google-research/mt-metrics-eval

To evaluate the accuracy of the generated error spans, we adopt the official protocol established in the WMT25 Metrics Task 2 (Lavie et al., 2025). We calculate the precision, recall, and micro-F1 scores at the character level to measure the overlap between the predicted and gold human error spans. This evaluation scheme rewards not only the exact character matches but additionally it incorporates a partial credit weighting of 0.5 points for cross-severity matches, meaning the model receives partial recognition if it correctly identifies the location of an error but misclassifies its severity.

It is important to note a limitation regarding the alignment of source-side errors within this specific evaluation framework. It accounts only for the errors that can be explicitly mapped to character indices within the target translation string. Consequently, it does not evaluate errors that can be annotated solely in the source sentence. A primary example is an omission error, where a phrase or an entire sentence is left untranslated and thus has no corresponding text span in the target output. For this reason, although our models are prompted to detect and extract these source-side omissions, we ignore them when calculating the F1 score. On the other hand, the omission errors are taken into account by the model when assigning the segment-level scalar score.

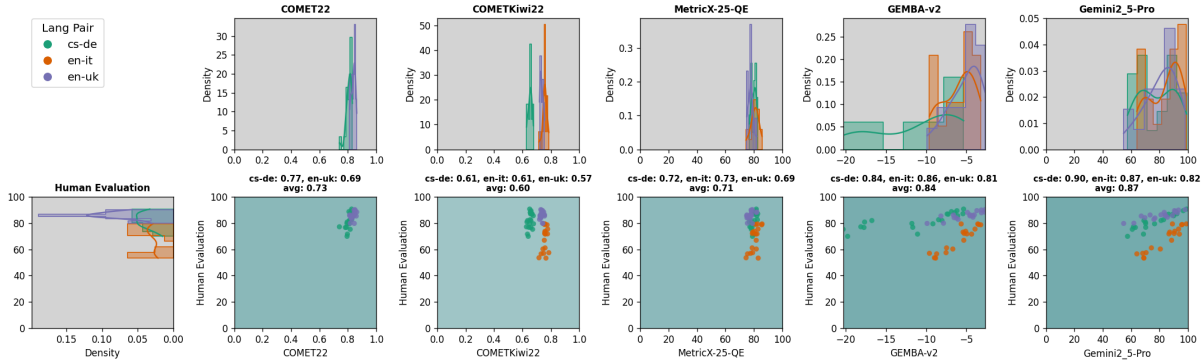

Figure 1: System-level performance of baseline metrics on the WMT25 dataset. The histograms display the score distributions for each metric, while the scatter plots illustrate the correlation between automated scores and human evaluations. Soft Pairwise Accuracy values for each language pair and their respective averages are provided above each plot. The background heatmap indicates the average accuracy across all language pairs (note: EN–IT is excluded for COMET22 due to the lack of available reference translations).

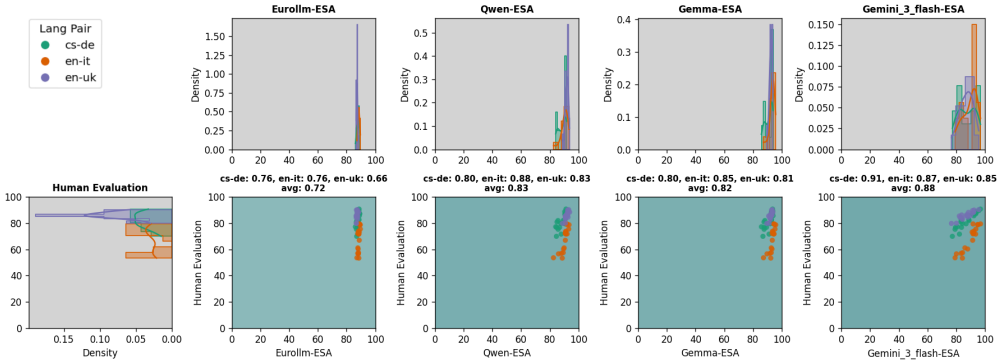

Figure 2: System-level performance of the proposed LLM-based QE metrics. The histograms display the distribution of scores generated by each model, while the scatter plots demonstrate the correlation with human-assigned ratings. Soft Pairwise Accuracy values for each language pair and their overall averages are indicated above each plot. The background heatmap represents the mean accuracy across all language pairs, with all proposed models providing quality scores on a scale of 0 to 100.

5 Results

5.1 Correlation with Human ESA

As illustrated in Figure 1 and Figure 2, the evaluated open-source models demonstrate highly competitive performance at the system level. Specifically, Qwen3 and Gemma achieve strong average SPA scores of 0.83 and 0.82, respectively. While these results are still lower than the proprietary Gemini-2.5-Pro (0.87) and GEMBA-MQM V2 (0.84), they substantially outperform traditional regression metrics such as COMET22 (0.73) and its reference-free counterpart COMETKiwi22 (0.60). Notably, these open-source models also surpass MetricX-25-QE (0.71), a SOTA model explicitly fine-tuned for the QE task. Among the open-weight models, EuroLLM exhibits a slightly lower average SPA of 0.72, though it remains competitive with the baselines.

Conversely, as seen in Figure 3 and Figure 4, the performance at the segment level reveals a persistent limitation of current open-source LLM evaluators. All evaluated open-source models exhibit significantly lower correlations compared to both proprietary LLMs and dedicated neural metrics. Only proprietary models achieve segment-level correlations that approach the reliability of traditional neural metrics. This aligns with previous findings that while LLMs excel at aggregating system-level quality, they struggle to consistently and reliably rank individual segments (Lavie et al., 2025). However, LLM-based evaluators can be viewed as complementary to these traditional metrics. While they may show lower ranking correlation, they provide a higher level of interpretability than regression-based models.

Furthermore, our evaluation highlights visible

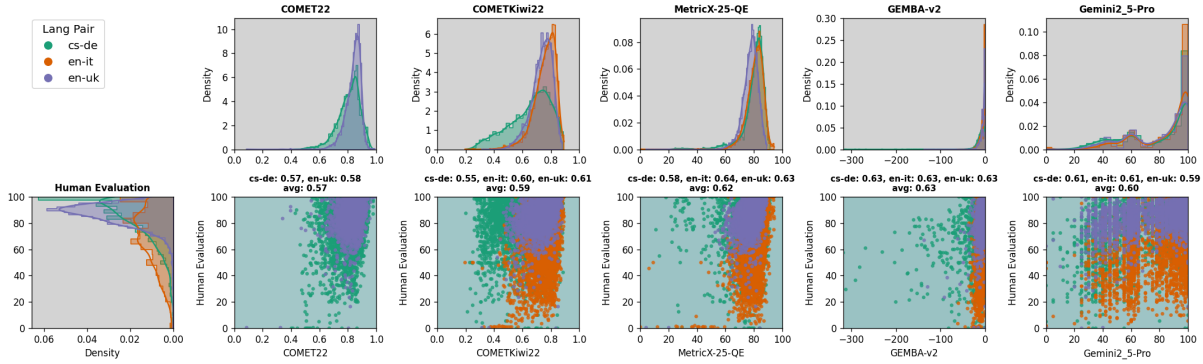

Figure 3: Segment-level performance of baseline metrics on the WMT25 dataset. The histograms show the distribution of scores for each metric across all evaluated segments. The scatter plots illustrate the correlation between automated metric scores and human judgments according to the "group-by-item" segment-level accuracy with tie calibration. Exact correlation values for each language pair and their respective averages are provided above each plot. The background heatmap represents the mean correlation across all language pairs (note: EN–IT is excluded for COMET22 due to a lack of available references).

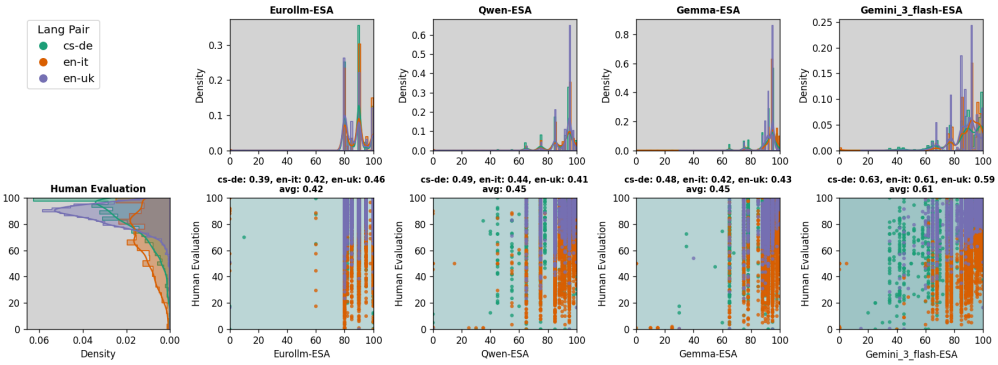

Figure 4: Segment-level performance of the proposed LLM-based QE metrics. The histograms display the distribution of quality scores generated by each model, while the scatter plots demonstrate the correlation with human-assigned ratings using the "group-by-item" segment-level accuracy with tie calibration. Pairwise correlation values for all language pairs and their overall averages are indicated above each plot. The background heatmap represents the mean correlation across all language pairs, with all proposed models providing scores on a scale of 0 to 100.

performance disparities across different language pairs. These variations are present for both the open-source and proprietary models. Because the model performance on specific language pairs differs depending on the pre-training data, carefully selecting and validating the appropriate model remains a crucial step for practical deployment.

Finally, an analysis of the score distributions reveals a pattern for generative models: score quantization. As illustrated in the scatter plots in Figure 4, LLM-generated scores frequently exhibit visible gaps, anchoring to specific numerical values. In contrast, traditional neural metrics yield smooth, continuous score distributions.

Additionally we present the results in tabular form in Appendix B.

5.2 Inter-Annotator Agreement

To contextualize the performance of the automated metrics, we computed the inter-annotator agreement between human evaluators who provided annotations for the WMT25 dataset. At the system level, presented in Figure 5, the average SPA between humans is only 0.71. Remarkably, all evaluated LLM-based QE metrics exceed this, demonstrating that automated metrics can provide more consistent system-level rankings than a secondary human annotator. Furthermore, human agreement exhibits substantial variance across language pairs, mirroring the observations made in automated metrics.

At the segment level, presented in Figure 6, the average human correlation drops significantly to 0.56. This confirms that segment-level scoring is an inherently subjective task, making it more dif-

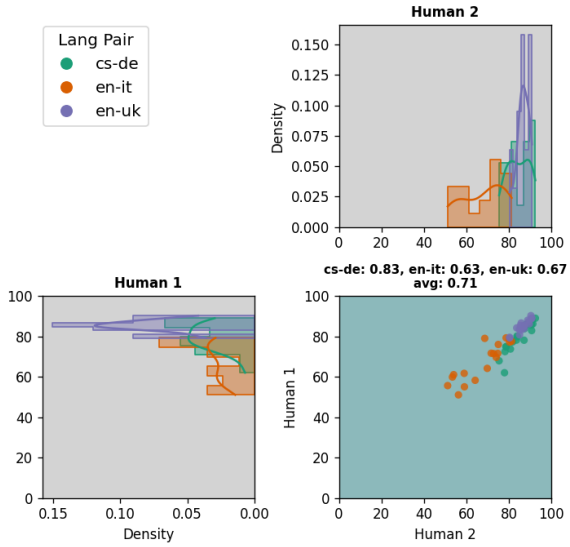

Figure 5: Inter-annotator agreement between independent human evaluators at the system level. The histograms (top and left) display the distribution of average quality scores assigned by each human annotator across all evaluated systems. The scatter plot (bottom-right) illustrates the correlation between the two sets of human scores. Soft Pairwise Accuracy values for each language pair and their overall average are provided above the plot.

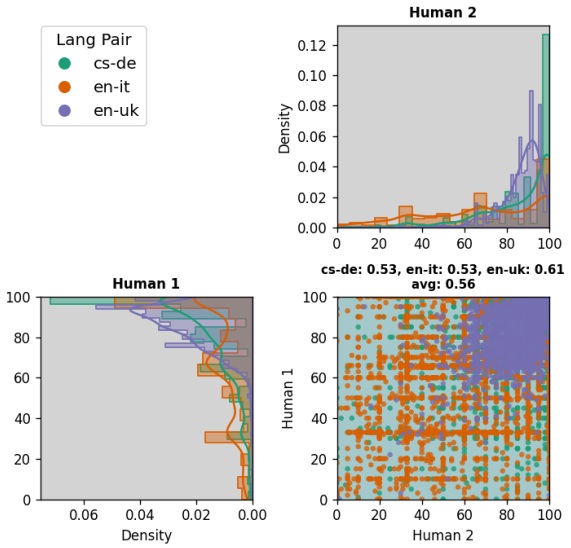

Figure 6: Inter-annotator agreement between independent human evaluators at the segment level. The histograms (top and left) show the distribution of quality scores assigned to individual segments by each human annotator. The scatter plot (bottom-right) demonstrates the correlation between these segment-level judgments. The "group-by-item" segment-level accuracy with tie calibration values for all language pairs and their overall average are provided above the plot.

ficult to model than system-level scoring. While the open-source models currently fall short in this task, Gemini actually achieves higher agreement with the gold standard.

Finally, the human-to-human comparison provides insight into the score quantization observed in the LLM outputs. The scatter plot in Figure 6 reveals that human evaluators also tend to anchor their assessments to specific round numbers, creating a grid-like visual pattern that is particularly pronounced in the English-Italian data. This indicates that the discretized scoring behavior of generative models is not strictly an artificial flaw, but might reflect a human tendency to categorize subjective quality into distinct, quantized tiers.

5.3 Error Span Detection

As shown in Table 6, the absolute F1 scores for error span detection remain consistently low across all models, reaching a maximum of approximately 18% for the proprietary Gemini-3-Flash and roughly 10% for the best open-source model, Gemma-3-27b-it. However, evaluating automated error span detection is complicated by inherently low human inter-annotator agreement. For the language pairs examined in our study, the micro-F1 agreement between independent human annotators ranges from only 30% to 37% (Lavie et al., 2025), which establishes a low empirical ceiling for the task.

Furthermore, by applying simple error filtering, we can effectively limit the number of annotated errors without any significant decrease in the overall F1 score. This filtering process makes the annotations more practically useful by curbing over-annotation. For instance, Qwen3 exhibits highest recall prior to filtering (e.g., 25.75 on English-Italian) but low precision (3.03), suggesting a strong tendency to over-annotate or hallucinate spans, which is confirmed by the amount of filtered errors presented in Table 2. After filtering, Qwen3 improves its precision without sacrificing much recall, increasing its F1 score. Overall, Gemma-3-27b-it proves to be the most capable open-source model for span detection, achieving the highest F1 scores, though this performance may simply be due to the model having the largest number of active parameters.

In addition to evaluating the span-detection metrics against human-annotated error spans, we also measured performance against the error spans generated by Gemini-3-Flash to determine if the open-source models exhibit higher alignment with a state-of-the-art proprietary LLM. As shown in Table 4, the open-source models reach a notably higher agreement with Gemini than with human annotators, demonstrating significant increases in

	Unfiltered errors			Filtered errors		
	Precision	Recall	F1 score	Precision	Recall	F1 score
Czech → German						
EuroLLM-9B-Instruct	6.21	9.03	7.36	6.59	8.49	7.42
Qwen3-VL-30B-A3B-Instruct	4.55	**20.39**	7.43	6.87	18.41	10.01
Gemma-3-27b-it	11.12	9.86	10.45	11.24	9.62	10.37
Gemini-3-Flash	**16.53**	20.23	**18.19**	**16.66**	**19.97**	**18.17**
English → Italian						
EuroLLM-9B-Instruct	4.35	7.43	5.49	4.44	7.16	5.48
Qwen3-VL-30B-A3B-Instruct	3.03	**25.75**	5.42	4.32	**23.42**	7.30
Gemma-3-27b-it	9.78	11.69	10.65	9.82	11.47	10.58
Gemini-3-Flash	**10.13**	22.10	**13.89**	**10.20**	21.96	**13.93**
English → Ukrainian						
EuroLLM-9B-Instruct	2.28	8.89	3.63	2.30	8.41	3.61
Qwen3-VL-30B-A3B-Instruct	1.51	23.43	2.84	1.84	20.65	3.39
Gemma-3-27b-it	3.77	11.37	5.67	3.76	11.11	5.61
Gemini-3-Flash	**4.57**	**24.72**	**7.72**	**4.58**	**24.27**	**7.71**

Table 3: Span-level error detection performance across three language pairs, evaluated against official human annotations. The table shows precision, recall, and F1 scores for four QE models, considering both unfiltered and filtered error sets. The highest score for each metric within a language pair is indicated in bold, and a heatmap is applied to the F1 score column.

	Unfiltered errors			Filtered errors		
	Precision	Recall	F1 score	Precision	Recall	F1 score
Czech → German						
EuroLLM-9B-Instruct	9.58	11.38	10.40	10.16	10.92	10.52
Qwen3-VL-30B-A3B-Instruct	8.11	**29.72**	12.74	12.26	**27.41**	16.95
Gemma-3-27b-it	**24.82**	17.98	**20.86**	**24.83**	17.73	**20.69**
English → Italian						
EuroLLM-9B-Instruct	12.30	9.64	10.81	12.54	9.40	10.75
Qwen3-VL-30B-A3B-Instruct	8.38	**32.69**	13.34	11.97	**30.11**	17.13
Gemma-3-27b-it	**27.04**	14.81	**19.14**	**27.10**	14.71	**19.07**
English → Ukrainian						
EuroLLM-9B-Instruct	12.87	9.30	10.80	13.17	9.10	10.76
Qwen3-VL-30B-A3B-Instruct	10.11	**28.98**	14.99	12.81	**27.07**	17.39
Gemma-3-27b-it	**27.60**	15.38	**19.75**	**27.43**	15.30	**19.64**

Table 4: Precision, recall, and F1 scores for span-level error detection, using Gemini-3-Flash predictions as the gold standard. This evaluation measures the alignment of open-source QE models with the proprietary `Gemini-3-Flash` model across three language pairs. Performance is reported for both unfiltered and filtered error sets, with the highest values per metric bolded and F1 scores visualized via a heatmap.

both precision and recall across all tested language pairs. Although this model-to-model agreement is still lower than the human inter-annotator agreement, it demonstrates that open-source models may be capable of approximating the proprietary LLMs.

6 Conclusion

In this work, we demonstrated the viability of using smaller, open-source LLMs for interpretable translation QE. By utilizing a streamlined prompting strategy, our approach efficiently generates quality scores, MQM-style error spans with suggested corrections, and full segment post-editions in a single pass. This unified approach reduces computational overhead and provides immediate, actionable feedback that is highly valuable for professional translation workflows.

Our evaluation revealed that open-source models are highly effective for system-level ranking. They consistently outperform traditional neural regression metrics, fine-tuned QE models, and even

the human inter-annotator agreement. However, segment-level and span-level QE remain a challenge, both for small open-source and larger proprietary LLMs. Our analysis demonstrated that these tasks are highly subjective, resulting in low inter-annotator agreement. To address the tendency of LLMs to over-annotate, we showed that a simple heuristic filtering pipeline can prove useful.

7 Future Work

Although our framework generates segment post-editions alongside quality scores and error annotations, we leave evaluating these corrections for future work. It would be beneficial to investigate how explicit error detection affects the quality of these post-editions and how they could be used to refine our filtering pipeline by verifying error spans against the model's suggested corrections to further reduce false positives.

Furthermore, the relationship between the generated error annotations and the final QE scores requires further analysis. Subsequent research could examine how the number and severity of identified errors impact the assigned score and its correlation with human judgment.

8 Limitations

The open-source models evaluated in this study are general-purpose multilingual LLMs operating in a few-shot setting. Because they were not explicitly fine-tuned for the QE task, their performance relies entirely on their pre-trained instruction-following capabilities. Task-specific fine-tuning on MQM or ESA datasets could potentially yield significant performance improvements and better alignment with human annotation logic.

Furthermore, to accommodate the VRAM constraints of consumer-grade GPUs, our approach evaluates translation segments in isolation. This omission of document-level context prevents the models from detecting broader discourse-level errors, such as cross-sentence coreference failures. On top of that, our evaluation framework currently does not incorporate external terminological specifications, cultural context, or details regarding the intended target audience. In professional localization workflows, these contextual layers are often critical for determining the true severity of an error or the overall appropriateness of a translation.

Additionally, the character-level F1 metric used for span-level error detection evaluation is strictly based on span overlap and severity weighting. It does not assess the accuracy of the generated error typology (e.g., distinguishing a grammatical error from a mistranslation) or the validity of the suggested corrections. Consequently, the current evaluation paradigm may underestimate the practical utility of the model's output by rewarding only the span of the detected error, while ignoring the usefulness of the provided feedback.

References

[Baccianella2025] Baccianella, Stefano. 2025. Json repair - a python module to repair invalid json, commonly used to parse the output of llms.

[Bai et al.2025] Bai, Shuai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. 2025. Qwen3-vl technical report.

[Deutsch et al.2023] Deutsch, Daniel, George Foster, and Markus Freitag. 2023. Ties matter: Meta-evaluating modern metrics with pairwise accuracy and tie calibration. In Bouamor, Houda, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12914–12929, Singapore, December. Association for Computational Linguistics.

[Fernandes et al.2023] Fernandes, Patrick, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In Koehn, Philipp, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 1066–1083, Singapore, December. Association for Computational Linguistics.

[Google DeepMind2025] Google DeepMind. 2025. Gemini 3 Flash model card. https://storage.googleapis.com/deepmind-media/Model-Cards/

Gemini-3-Flash-Model-Card.pdf, December. Accessed: 2026-03-13.

[Grattafiori et al.2024] Grattafiori, Aaron, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena,

Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models.

[Guerreiro et al.2024] Guerreiro, Nuno M., Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

[Jiang et al.2024] Jiang, Albert Q., Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

[Junczys-Dowmunt2025] Junczys-Dowmunt, Marcin. 2025. GEMBA v2: Ten judgments are better than one. In Haddow, Barry, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Tenth Conference on Machine Translation*, pages 926–933, Suzhou, China, November. Association for Computational Linguistics.

[Juraska et al.2025] Juraska, Juraj, Tobias Domhan, Mara Finkelstein, Tetsuji Nakagawa, Geza Kovacs, Daniel Deutsch, Pidong Wang, and Markus Freitag. 2025. MetricX-25 and GemSpanEval: Google Translate submissions to the WMT25 evaluation shared task. In Haddow, Barry, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Tenth Conference on Machine Translation*, pages 957–968, Suzhou, China, November. Association for Computational Linguistics.

[Kocmi and Federmann2023a] Kocmi, Tom and Christian Federmann. 2023a. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In Koehn, Philipp, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore, December. Association for Computational Linguistics.

[Kocmi and Federmann2023b] Kocmi, Tom and Christian Federmann. 2023b. Large language models are state-of-the-art evaluators of translation quality. In Nurminen, Mary, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, Eva Vanmassenhove, Sergi Alvarez Vidal, Nora Aranberri, Mara Nunziatini, Carla Parra Escartín, Mikel Forcada, Maja Popovic, Carolina Scarton, and Helena Moniz, editors, *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland, June. European Association for Machine Translation.

[Kocmi et al.2024] Kocmi, Tom, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024. Error span annotation: A balanced approach for human evaluation of machine translation. In Haddow, Barry, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453, Miami, Florida, USA, November. Association for Computational Linguistics.

[Lavie et al.2025] Lavie, Alon, Greg Hanneman, Sweta Agrawal, Diptesh Kanojia, Chi-Kiu Lo, Vilém

Zouhar, Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, Archchana Sindhujan, Jiayi Wang, David Ifeoluwa Adelani, Brian Thompson, Tom Kocmi, Markus Freitag, and Daniel Deutsch. 2025. Findings of the WMT25 shared task on automated translation evaluation systems: Linguistic diversity is challenging and references still help. In Haddow, Barry, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Tenth Conference on Machine Translation*, pages 436–483, Suzhou, China, November. Association for Computational Linguistics.

[Lommel et al.2013] Lommel, Arle Richard, 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, November 28-29. Aslib.

[Lu et al.2024] Lu, Qingyu, Baopu Qiu, Liang Ding, Kanjian Zhang, Tom Kocmi, and Dacheng Tao. 2024. Error analysis prompting enables human-like translation evaluation in large language models. In Ku, Lun-Wei, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8801–8816, Bangkok, Thailand, August. Association for Computational Linguistics.

[Lu et al.2025] Lu, Qingyu, Liang Ding, Kanjian Zhang, Jinxia Zhang, and Dacheng Tao. 2025. MQM-APE: Toward high-quality error annotation predictors with automatic post-editing in LLM translation evaluators. In Rambow, Owen, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5570–5587, Abu Dhabi, UAE, January. Association for Computational Linguistics.

[Martins et al.2025] Martins, Pedro Henrique, João Alves, Patrick Fernandes, Nuno M. Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M. Alves, José Pombal, Nicolas Boizard, Manuel Faysse, Pierre Colombo, François Yvon, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2025. Eurollm-9b: Technical report.

[Rei et al.2020] Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In Webber, Bonnie, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online, November. Association for Computational Linguistics.

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

[Team et al.2025] Team, Gemma, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob

Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 technical report.

[Thompson et al.2024] Thompson, Brian, Nitika Mathur, Daniel Deutsch, and Huda Khayrallah. 2024. Improving statistical significance in human evaluation of automatic metrics via soft pairwise accuracy. In Haddow, Barry, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 1222–1234, Miami, Florida, USA, November. Association for Computational Linguistics.

A Prompts

```
You are an AI assistant specialized in {source_language}-to-{target_language}
translation quality assurance.

You will receive a pair of paragraphs as a JSON structure. For each input, reply
with an extended JSON object that contains the information described below.

First, rate the quality of the translation on a scale from 1 to 100, taking the
source text into account. If the translation is fully accurate and consistent with
the source sentence, return 100. If the translation is not fully accurate and
consistent with the source sentence, return a score between 1 and 100 depending on
the quality of the translation and the errors present in the translation. If the
translation is in a different language than requested, return 0.
Then post-edit the translation. Correct any errors or unnatural phrasing, ensuring
accuracy, fluency, and fidelity to the source text. If the original translation is
already correct, leave it unchanged.
Then focus on particular errors in the original translation, that you have
corrected in the post-edited translation. Each error is classified as one of three
categories: "critical", "major", and "minor". Critical errors inhibit comprehension
of the text, change the meaning, provide false information, or make the text
impossible to understand. Major errors disrupt the flow, make the text difficult to
read or awkward, but the original meaning is still recoverable and understandable.
Minor errors are technically incorrect, such as typos or punctuation, but do not
affect the meaning, readability, or flow of the translated text. For every of the
main three categories, additionally identify error types in the translation and
sub-classify them. The types of errors are: "accuracy" ("addition",
"mistranslation", "omission", "untranslated text"), "fluency" ("character
encoding", "grammar", "inconsistency", "punctuation", "register", "spelling"),
"style" ("awkward"), "terminology" ("inappropriate for context", "inconsistent
use"), "non-translation", or "other". For every error type, supply suggested
correction of the error ("correction") and a very short and concise description of
the error ("short_desc"). The correction should be a single suggested word or
phrase that is a direct replacement for the error. If there are no errors of a
specific main category (critical, major or minor), it is OK to return an empty list
for that category. It also OK to not return any errors for any category if
everything is fine.

Here is an example of a JSON input for English-German translation:
{
   "source_language": "English",
   "source": "I do apologise about this, we must gain permission from the account
   holder to discuss an order with another person, I apologise if this was done
   previously, however, I would not be able to discuss this with yourself without
   the account holders permission.",
   "translation_language": "German",
   "translation": "Ich entschuldige mich dafür, wir müssen die Erlaubnis einholen,
   um eine Bestellung mit einer anderen Person zu involvment. Ich entschuldige
   mich, falls dies zuvor geschehen wäre, aber ohne die Erlaubnis des Kontoinhabers
   wäre ich nicht in der Lage, dies mit dir involvement."
}

And here is a corresponding JSON output with the score, post-edited translation and
error annotations:
{
   "score": 82,
   "post_edited_translation": "Ich entschuldige mich dafür, wir müssen die
   Erlaubnis einholen, um eine Bestellung mit Kontoinhaber zu besprechen. Ich
   entschuldige mich, falls dies zuvor geschehen kann, aber ohne die Erlaubnis des
   Kontoinhabers kann ich nicht in der Lage, dies mit Sie involvement.",
   "errors": {
      "critical": [],
      "major": [
         {"type": "accuracy/mistranslation", "source_error": "discuss",
         "target_error": "involvment", "correction": "besprechen", "short_desc":
         "'discuss' is mistranslated as 'involvment'"},
         {"type": "accuracy/omission", "source_error": "the account holder",
         "target_error": None, "correction": "Kontoinhaber", "short_desc": "'the
         account holder' is missing"}
      ],
```

```
      "minor": [
         {"type": "fluency/grammar", "source_error": None, "target_error": "wäre",
         "correction": "kann", "short_desc": "'wäre' is a bit awkward"},
         {"type": "fluency/register", "source_error": None, "target_error": "dir",
         "correction": "Sie", "short_desc": "'dir' should be 'Sie'"}
      ]
   },
}

Here is an example of a JSON input for Polish-Spanish translation:
{
   "source_language": "Polish",
   "source": "Szanowny Kliencie, informujemy, że Twoja przesyłka została dzisiaj
   nadana i powinna dotrzeć w ciągu 3 dni roboczych. Prosimy o sprawdzenie skrzynki
   mailowej w celu uzyskania numeru śledzenia.",
   "translation_language": "Spanish",
   "translation": "Estimado Cliente, informamos que tu envío fue hoy sobre y debe
   llegar en tres días trabajan. Por favor, chequea el mail para el número de
   rastrear."
}

And here is a corresponding JSON output with the score, post-edited translation and
error annotations:
{
   "score": 45,
   "post_edited_translation": "Estimado Cliente, le informamos que su paquete fue
   enviado hoy y debería llegar en un plazo de 3 días hábiles. Por favor, revise su
   correo electrónico para obtener el número de seguimiento.",
   "errors": {
      "critical": [
         {"type": "accuracy/omission", "source_error": "przesyłka", "target_error":
         None, "correction": "paquete/envío", "short_desc": "Meaning is changed by
         omitting 'package' or 'shipment'"},
         {"type": "accuracy/mistranslation", "source_error": "nadana",
         "target_error": "sobre", "correction": "enviado", "short_desc": "'nadana'
         is incorrectly translated as 'sobre' (envelope/about)"}
      ],
      "major": [
         {"type": "fluency/grammar", "source_error": "roboczych", "target_error":
         "trabajan", "correction": "hábiles", "short_desc": "'trabajan' is
         incorrect word form, should be 'hábiles'"},
         {"type": "fluency/grammar", "source_error": "w ciągu", "target_error":
         "en", "correction": "en un plazo de", "short_desc": "Preposition 'en' or
         'en un plazo de' is missing"},
         {"type": "style/awkwardness", "source_error": "sprawdzenie",
         "target_error": "chequea", "correction": "revise", "short_desc":
         "'chequea' is too informal, 'revise' is better for formal communication"}
      ],
      "minor": [
         {"type": "fluency/grammar", "source_error": "skrzynki mailowej",
         "target_error": "el mail", "correction": "el correo electrónico",
         "short_desc": "Missing article 'el' before 'mail'"},
         {"type": "fluency/grammar", "source_error": "śledzenia", "target_error":
         "rastrear", "correction": "seguimiento", "short_desc": "Gerund 'rastrear'
         is wrong, should be noun 'seguimiento'"}
      ]
   }
}

Here is an example of a JSON input for French-Italian translation:
{
   "source_language": "French",
   "source": "En raison d'une forte demande, il peut y avoir un léger retard dans
   le traitement de votre commande. Nous vous remercions de votre patience et de
   votre compréhension.",
   "translation_language": "Italian",
   "translation": "A causa di una forte domanda, potrebbe esserci un leggero
   ritardo nell'elaborazione del vostro ordine. Vi ringraziamo per la vostra
   pazienza e la vostra comprensione."
```

```
}

And here is a corresponding JSON output with the score, post-edited translation and
error annotations:
{
   "score": 100,
   "post_edited_translation": "A causa di una forte domanda, potrebbe esserci un
   leggero ritardo nell'elaborazione del vostro ordine. Vi ringraziamo per la
   vostra pazienza e la vostra comprensione.",
   "errors": {
      "critical": [],
      "major": [],
      "minor": []
   }
}

If the whole translation is in the wrong language, return one "untranslated text"
critical error for the whole sentence and no other errors.
IMPORTANT: All annotated errors and assigned ratings should refer only to the
translation and not to the source text.

You will receive the {source_language}-{target_language} translation pair as a JSON
object. Analyze the translation as discussed above and produce a JSON object with
the analysis in response. Do not invent structural elements that are not present in
the JSON examples above. The only allowed keys are "score",
"post_edited_translation", "errors", "critical", "major", "minor", "type",
"short_desc", "source_error", "target_error", "correction".
```

Listing 1: QE system prompt.

```
{
   "source_language": {source_language},
   "source": {source_text},
   "translation_language": {target_language},
   "translation": {target_text}
}
```

Listing 2: QE user prompt.

B Results

COMET22	COMETKiwi22	MetricX-25-QE	GEMBA-v2	Gemini2_5-Pro	Eurollm-ESA	Qwen-ESA	Gemma-ESA	Gemini_3_flash-ESA
Czech → German								
0.77	0.61	0.72	0.84	0.90	0.76	0.80	0.80	**0.91**
English → Italian								
-	0.61	0.73	0.86	0.87	0.76	**0.88**	0.85	0.87
English → Ukrainian								
0.69	0.57	0.69	0.81	0.82	0.66	0.83	0.81	**0.85**
Average								
0.73	0.60	0.71	0.84	0.87	0.72	0.83	0.82	**0.88**

Table 5: System-level correlation between automated scores (both baselines and the proposed LLM-based QE metrics) and human evaluations for the WMT25 dataset, expressed as Soft Pairwise Accuracy. The table provides a comparative analysis for three specific language pairs and the average performance across all evaluated directions. In each row, the highest accuracy value is indicated in bold to highlight the most effective metric.

COMET22	COMETKiwi22	MetricX-25-QE	GEMBA-v2	Gemini2_5-Pro	Eurollm-ESA	Qwen-ESA	Gemma-ESA	Gemini_3_flash-ESA
Czech → German								
0.57	0.55	0.58	**0.63**	0.61	0.39	0.49	0.48	**0.63**
English → Italian								
-	0.60	**0.64**	0.63	0.61	0.42	0.44	0.42	0.61
English → Ukrainian								
0.58	0.61	**0.63**	**0.63**	0.59	0.46	0.41	0.43	0.59
Average								
0.57	0.59	0.62	**0.63**	0.60	0.42	0.45	0.45	0.61

Table 6: Segment-level correlation between automated scores (both baselines and the proposed LLM-based QE metrics) and human judgments on the WMT25 dataset. The evaluation utilizes the "group-by-item" accuracy with tie calibration to measure how well each metric aligns with human ratings at the segment level. The table details results for three language pairs along with their overall average. For each row, the maximum correlation value is highlighted in bold.