NVIDIA-NeMo's WMT 2025 Metrics Shared Task Submission

Brian Yan[1], **Shuoyang Ding**[2], **Kuang-Da Wang**[3],
Siqi Ouyang[1], **Oleksii Hrinchuk**[2], **Vitaly Lavrukhin**[2], **Boris Ginsburg**[2]
[1]Carnegie Mellon University, [2]NVIDIA, [3]National Yang Ming Chiao Tung University
byan@cs.cmu.edu, shuoyangd@nvidia.com

Abstract

This paper describes NVIDIA-NeMo's WMT 2025 Metrics Shared Task submission. We investigated two strategies for extending Machine Translation (MT) evaluation to unsegmented documents: 1) first segmenting into sentences and then applying regression-based metrics on aligned sentence pairs, and 2) directly utilizing the long-context capabilities of LLMs. The base comparison of the segmentation-based and LLM-based metrics on the WMT 2023-24 evaluation sets indicated that the former performs more robustly across language pairs. Thus we sought to improve the LLM-based approach by incorporating *relative evaluation* - this setting jointly evaluates all candidate translations at once and relative to each other, rather than evaluating each separately. Our experiments using the open-source Qwen3 LLM show that relative evaluation improves score correlations with human judgment, but only if the task is structured as a 2-stage evaluate-then-refine problem.

1 Introduction

For most of its history, machine translation (MT) research has revolved around the sentence as the primary unit of translation and evaluation. Standard MT benchmarks and evaluation protocols typically present systems with short, isolated segments, and assess their quality against human-produced references (Kocmi et al., 2024). This sentence-level paradigm has been driven by practical constraints: statistical and neural MT systems were long limited by both computational costs and modeling power (Kim et al., 2019). And in turn, automatic metrics, whether lexical (Papineni et al., 2002) or regression-based (Rei et al., 2020), were also designed to operate on sentence-like segments.

However, translation in real-world applications rarely occurs in isolation. Professional translators work with continuous documents, where meaning and style are shaped by discourse, context, and document-level coherence (Lommel et al., 2014). As neural MT systems have become more capable, particularly with the advent of large language models (LLMs), research has increasingly shifted toward document-level translation and evaluation (Zhu et al., 2024; Pang et al., 2025). This shift reflects both a practical need for evaluating translations as they appear in natural, unsegmented contexts and a growing recognition that many important aspects of quality, such as pronoun resolution, lexical consistency, and narrative flow, can only be judged when viewing the text holistically. This recent emergence of long-context MT models has made it technically feasible to translate much larger spans of text at once, raising a new research question in the field of MT evaluation: *is it better to apply existing sentence-level metrics to segmented text or apply LLM-based metrics directly to document-length text?*

In the WMT 2025 Metrics Shared Task, we focus specifically on the problem of unsegmented document evaluation: assessing the quality of MT outputs when the entire document is presented as a single sequence. Our submission compares two divergent strategies: (1) applying traditional regression-based metrics after segmenting documents into sentences, and (2) leveraging the long-context capabilities of LLMs to perform holistic document evaluation directly. To support our investigation, we simulated document-level human evaluations by concatenating the sentence-level MQM annotations from the WMT 2023 and 2024 Metrics Shared Tasks. Our experiments showed that the scaled up regression-based metric correlated better with human judgement than the LLM-based metric – we thus designed the former as our primary submission and the latter as our secondary submission to the WMT 2025 Metrics Shared Task.

While our primary findings favored the segmentation-based regression metric, we also explored ways to strengthen the LLM-based approach.

Proceedings of the Tenth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 920–925
November 8–9, 2025.

In particular, we experimented with *relative evaluation* — a setup in which all candidate translations for a given source are presented to the model simultaneously and ranked against each other, rather than scored in isolation. We found that relative evaluation yielded only marginal improvements in correlation with human judgments, and only when implemented as a two-stage "evaluate-then-refine" process. Although these gains were not large enough to change our overall ranking of the two strategies, this line of work remains relevant for understanding how LLM prompting strategies interact with long-context MT evaluation. We therefore complement our main results with an analysis of why relative evaluation showed limited benefits and where it may still hold promise.

2 Document-level Human Evaluation

Prior to 2025, the WMT shared tasks were limited to the segment-level translation paradigm - all MT systems and all human evaluation operated on segments. To study document-level evaluation, we construct a simulated document-level MQM set by concatenating the source and target segments of each document and summing their MQM error counts. Table 1 compares the source lengths and MQM scores before and after document concatenation.

While this offers the best available proxy for document-level score correlations on WMT data, it has two notable limitations:

1. MT systems trained and evaluated on segmented inputs may differ from true document-level MT systems, which can exhibit distinct error patterns such as over- or under-translation.

2. MQM raters were instructed to consider surrounding context, but their judgments remained segment-focused; discourse-level phenomena may be underreported.

Note that we do not simulate document-level evaluation for the WMT 2024 sets because a large portion of segments were not evaluated. We also do not simulate document-level ESA data, as its 0–100 scale is not naturally additive, unlike MQM error counts. We therefore focus on the three language pairs from WMT 2023: En-De, He-En, and Zh-En.

Table 1: Comparison of segment and document-level WMT human evaluation data: average source character length (Len) and average MQM score (Score).

	Segment		Document	
Set	Len	Score	Len	Score
WMT23-En-De	354	-7.1	1028	-16.9
WMT23-He-En	84	-2.3	1719	-17.3
WMT23-Zh-En	40	-4.3	449	-25.1
WMT24-En-De	185	-3.0	-	-
WMT24-En-Es	185	-1.0	-	-
WMT24-Ja-Zh	91	-3.2	-	-

3 Metric Descriptions

3.1 Segmentation-based

The segmentation-based system breaks down document-level evaluation into sentence-level subproblems and relies on legacy sentence-level metrics for evaluation. Our approach first segments the source and target documents into sentences using ersatz (Wicks and Post, 2021), then establish aligned sentence blocks from these sentences using Vecalign (Thompson and Koehn, 2019) and LASER (Artetxe and Schwenk, 2019) sentence embeddings. Because our preliminary finding shows that existing sentence-level metrics are not good at identifying over- and under-translation errors, we have to rely on null alignments to identify these errors. To avoid merging unrelated sentences into aligned sentence blocks, we introduce an adaptive heuristic search strategy that dynamically finds the optimal alignment penalty (β_{skip}) for each document, ensuring those over- and under-translation errors are correctly identified as null alignments. Different from past practices such as mWERSegmenter (Matusov et al., 2005), which jointly align and segment system translations according to a reference, our system is a reference-free submission that directly aligns system translation to the source.

We apply MetricX-24-XXL-Hybrid-QE (Juraska et al., 2024) to each aligned blocks. In cases where null alignments are established, we penalize null alignments by assigning a score of 25 for each null alignment. We average the scores evalauted over sentence blocks to obtain a document-level score. To ensure the score aligns with the scale and directionality of WMT 2025, we apply a simple linear transformation of $100 - 4 \times s_d$ on the document-level score s_d.

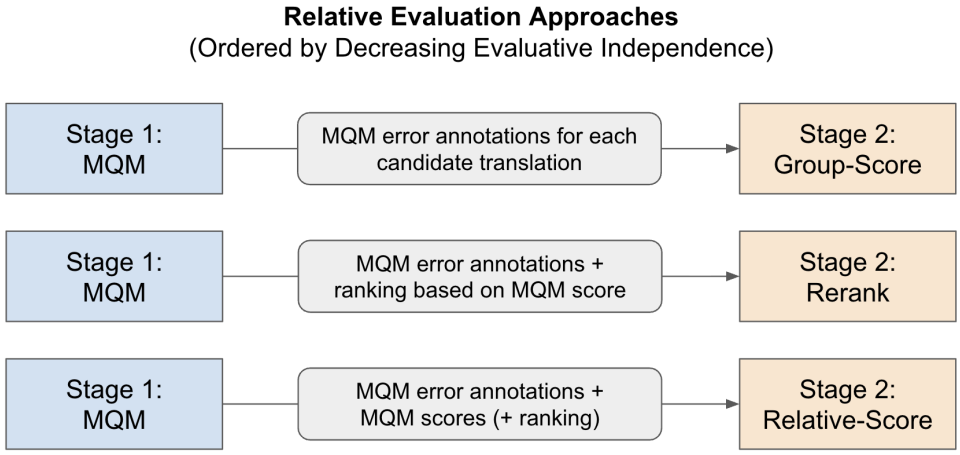

Figure 1: Overview of our 2-stage LLM-based methods: three relative evaluation approaches incorporating varying degrees of information from the initial MQM stage.

3.2 LLM-based

We implemented four document-level MT evaluation methods based on the Qwen3 large language model (LLM). All methods operate on unsegmented documents and follow the MQM (Multidimensional Quality Metrics) annotation scheme. The first method is a direct re-implementation of an established approach, while the remaining three extend it by incorporating a second-stage relative evaluation procedure.

3.2.1 Qwen-MQM

The baseline method, Qwen-MQM, is a Qwen-based re-implementation of the GEMBA-MQM framework, which was originally proposed using GPT models (Zhao et al., 2024). In this setting, the LLM is provided with the source document and a candidate translation and is prompted to produce MQM annotations (Kocmi and Federmann, 2023). The final document score is obtained by aggregating these annotations according to the MQM weighting scheme (Freitag et al., 2021). This method evaluates each system output independently, without reference to other candidate translations.

3.2.2 Relative Evaluation Extensions

The remaining three methods augment Qwen-MQM with a second evaluation stage in which the LLM is shown multiple system outputs for the same source document and instructed to re-assess or adjust scores based on cross-system comparison. This relative evaluation setting leverages the LLM's long-context capability to consider multiple translations simultaneously, with the goal of improving the utility of the metric for ranking MT systems. We present the methods in order of decreasing evaluative independence - that is, the degree to which final scoring decisions are formed without being conditioned on the scores and rankings of the initial MQM stage.

Qwen-MQM-Group-Score: In the second stage, the LLM is presented with the complete set of candidate translations together with their initial MQM annotations. It is then instructed to assign final scores to all candidates in a single pass, explicitly weighing the relative strengths and weaknesses that emerge through comparison. Since only the initial MQM annotations, and not the resultant MQM scores, are presented to the LLM in this second stage we consider that this approach has a high degree of evaluative independence.

Qwen-MQM-Rerank: In the second stage, the LLM is shown the initial ranking of candidates derived from their MQM scores, along with their MQM annotations, and is asked whether any adjustments to the ordering are warranted. Compared to the previous approach, this one offers less evaluative independence in the second stage: here, the initial ranking is presented to the LLM and the task is framed as a re-ranking of the initial judgments.

Qwen-MQM-Relative-Score: In the second stage, the LLM receives the initial MQM annotations and MQM scores (and thus implicitly the initial rankings) of all other candidates and is tasked with assigning a score to a single target translation in light of these scores. This creates the strongest dependency on prior judgments, as the evaluation of each candidate is explicitly anchored to the assessed quality of its competitors. Unlike the previ-

Table 2: Comparison of regression-based and LLM-based metrics on segment-level WMT data, as measured by system and segment score correlations with human judgments (Kendall's Tau). Correlations are averaged across language pairs.

Model	Sys	Seg
XComet-QE	0.731	0.371
MetricX-QE	0.769	**0.387**
GEMBA-MQM/ESA	**0.809**	0.362
Llama-MQM	0.674	0.230
Qwen-MQM	0.736	0.356

ous two approaches which evaluate all candidates at once, this one cycles through each candidate in order to evaluate each relative to the rest.

Figure 1 summarizes these three relative evaluation approaches in terms of what information from the initial MQM stage is made available. In other words, with more information (in the form of annotations, ranking, and scores), the evaluative independence of the second stage decreases.

4 Experimental Setup

Data: We use evaluation data from the WMT 2023–2024 Metrics Shared Tasks: MQM (both years) and ESA (2024). These are inherently segment-level corpora: translations were produced from pre-segmented sources, and human ratings were applied to individual segments.

Models: For LLM-based metrics, we use Qwen3-14B (Team, 2025b) in our experiments. The Qwen3 series is a hybrid reasoning/instruction-following model, capable of scaling inference-time compute via the generation of reasoning traces; however, we found that the reasoning mode degraded performance, so we disabled it in our experiments. In our attempt to identify the best open-source LLM, we also experimented with Llama3.1-8B (Grattafiori et al., 2024) and Gemma-3-12b-it (Team, 2025a) – these were less performant than Qwen3, as described in the next section.

Meta-Evaluation: For measuring score correlation with human judgment we rely on Kendall's Tau, as computed by the official WMT repository: `https://github.com/google-research/mt-metrics-eval`. We track both system-level and segment/document-level score correlations.

5 Results

5.1 Segment-level Evaluation

Table 2 compares regression-based and LLM-based metrics on the WMT 2023–2024 segment-level datasets. Correlations with human judgments are reported at both the system and segment level, averaged over all language pairs.

MetricX-QE stands out as the best overall metric, considering both system and segment-level score correlation. The GPT-based GEMBA metric is a close second, but for our submission we opted for open-source alternatives. Therefore on the LLM side, Qwen-MQM is the strongest available metric. We found that Llama produced reasonable, but weaker, results while Gemma failed to consistently follow the MQM instructions.

5.2 Document-level Evaluation

Given that MetricX-QE and Qwen-MQM were the strongest regression-based and LLM-based metrics respectively at the segment level, we centered our document level investigations around these two.

Table 3 presents results on the simulated document-level dataset constructed from WMT 2023 MQM annotations. Here the LLM-base metric outperformed the regression-based, unlike in the previous segment-level setting. This suggests that the long context capability of LLMs lead to a more holistic evaluation of document-level translations, while regression-based methods still require some form of segmentation into parts.

The single stage Qwen-MQM outperformed the two relative evaluation approaches with the highest degrees of evaluative independence in the second stage: Qwen-MQM-Rerank and Qwen-MQM-Group-Score. These degradations resulting from the second stage suggest that the relative evaluation ability of LLMs is still weak.

On the other hand, the Qwen-MQM-Relative-Score approach yielded moderate improvements over Qwen-MQM - since this approach provides a great deal of information (MQM annotations and MQM scores) from the first stage, it limits how much the scores produced in the second stage can deviate from that of the first stage.

6 Conclusion

We investigated two strategies for unsegmented document-level MT evaluation: scaling traditional regression-based metrics to longer contexts and applying LLM-based metrics capable of holistic

Table 3: Comparison of regression-based and LLM-based metrics on document-level WMT data, as measured by system and document score correlations with human judgments (Kendall's Tau).

Metric	En-De		He-En		Zh-En		Avg	
	Sys	Doc	Sys	Doc	Sys	Doc	Sys	Doc
LASER-MetricX-QE	0.909	0.328	0.848	0.233	0.714	0.285	0.824	0.282
Qwen-MQM	0.909	**0.429**	0.758	0.346	0.810	0.460	0.836	**0.425**
Qwen-MQM-Rerank	0.909	0.415	0.758	0.343	0.829	0.460	0.832	0.406
Qwen-MQM-Group-Score	**0.939**	0.365	**0.909**	0.250	**0.924**	0.261	**0.924**	0.292
Qwen-MQM-Relative-Score	**0.939**	0.421	**0.909**	**0.357**	0.810	**0.471**	0.886	0.416

assessment. While regression-based metrics exhibited stronger correlations with human judgments than LLM-based metrics on segment-level data, the inverse was true on simulated document-level data. While relative evaluation techniques modestly improved LLM-based performance, gains were only achieved under fairly restrictive settings. These findings suggest that long-context LLMs are a promising basis for document-level MT evaluation, but further work is needed to fully realize the potential of LLM-based approaches.

References

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Yunsu Kim, Duc Thanh Tran, and Hermann Ney. 2019. When and why is document-level context useful in neural machine translation? In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 24–34.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, and 1 others. 2024. Findings of the wmt24 general machine translation shared task: The llm era is here but mt is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46.

Tom Kocmi and Christian Federmann. 2023. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica*, (12):0455–463.

Evgeny Matusov, Gregor Leusch, Oliver Bender, and Hermann Ney. 2005. Evaluating machine translation output with automatic sentence segmentation. In *2005 International Workshop on Spoken Language Translation, IWSLT 2005, Pittsburgh, PA, USA, October 24-25, 2005*, pages 138–144. ISCA.

Jianhui Pang, Fanghua Ye, Derek Fai Wong, Dian Yu, Shuming Shi, Zhaopeng Tu, and Longyue Wang. 2025. Salute the classic: Revisiting challenges of machine translation in the age of large language models. *Transactions of the Association for Computational Linguistics*, 13:73–95.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Gemma Team. 2025a. Gemma 3.

Qwen Team. 2025b. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1342–1348, Hong Kong, China. Association for Computational Linguistics.

Rachel Wicks and Matt Post. 2021. A unified approach to sentence segmentation of punctuated text in many languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3995–4007, Online. Association for Computational Linguistics.

Justin Zhao, Timothy Wang, Wael Abid, Geoffrey Angus, Arnav Garg, Jeffery Kinnison, Alex Sherstinsky, Piero Molino, Travis Addair, and Devvret Rishi. 2024. Lora land: 310 fine-tuned llms that rival gpt-4, A technical report. *CoRR*, abs/2405.00732.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.