Exploring The Impact of Document Context on LLM-based Machine Translation Evaluation

Anonymous ACL submission

Abstract

Human annotators typically assess translation quality by considering the broader document context, which enables them to detect errors arising from discourse-level phenomena that would be missed when evaluating text segments in isolation. However, the role of document-level context in automatic Machine Translation (MT) evaluation remains underexplored. In this work, we systematically investigate the impact of document context on LLM-based machine translation evaluation and its interaction with the availability of manually curated references. We find that the effect of document context varies substantially across the two test sets examined. To better understand this variation, we evaluate LLMs across several configurations, also exploring how performance varies when enabling the thinking inference mode and conducting task-specific fine-tuning. Our results show that fine-tuning improves performance on both test sets, with larger gains when document context is available. This suggests that base models may struggle with document-level evaluation, whereas fine-tuning can help them learn to leverage document context effectively.

1 Introduction

Machine Translation (MT) evaluation involves assessing the quality of translated text, and modern automatic evaluators (hereafter, auto-evaluators) are machine learning models designed to automate this task.

Traditionally, most MT evaluation approaches have focused on segment-level evaluation,[1] where segments of text are evaluated in isolation, independently of surrounding context (Rei et al., 2020, 2022; Kocmi and Federmann, 2023b,a; Juraska et al., 2023, 2024; Guerreiro et al., 2024). However, segment-level evaluation is limited because it prevents auto-evaluators from detecting errors arising from discourse-level phenomena, such as anaphora, i.e., the use of pronouns or expressions that refer to earlier words; lexical cohesion, i.e., terminology, synonyms, hyponyms, etc.; or ellipsis, i.e., the omission of words that can be understood from context (Wong and Kit, 2012; Bawden et al., 2018).

In human evaluation, Läubli et al. (2018) showed that document-level evaluation produces more accurate assessments of translation quality than segment-level. Following their study, subsequent editions of the Conference on Machine Translation (WMT) have adopted document-level human evaluation protocols when collecting annotations of translation quality (Barrault et al., 2019, 2020; Akhbardeh et al., 2021; Kocmi et al., 2022, 2023, 2024a, 2025).

In automatic evaluation, however, most approaches still operate at the segment level. Despite the lack of document-level context being mitigated by the presence of reference translations (indeed, many discourse-level phenomena can be disambiguated by consulting a ground-truth translation), reliance on reference translations may bias evaluation toward translations that are similar to the provided references, potentially compromising evaluation reliability, particularly when human-curated references are of low quality (Freitag et al., 2023). Furthermore, reference-based auto-evaluators have a very limited range of application scenarios, mostly concerning the evaluation of MT systems on benchmarks for which references have been made available in advance. By contrast, reference-free auto-evaluators have a much wider range of applications, including data filtering and translation re-ranking (Freitag et al., 2022a; Fernandes et al., 2022; Peter et al., 2023; Alves et al., 2024; Ramos et al., 2024; Perrella et al., 2024), and are also used as objectives for reinforcement learning-based fine-tuning of MT systems (Ramos et al., 2024; Xu et al., 2024; Zheng et al., 2025).

[1] A segment may be a sentence or a multi-sentence unit of text into which a longer document has been segmented.

This work investigates the effects of incorporating document context into LLM-based MT evaluation, and its relationship with reference translations. Specifically, we benchmark thinking and non-thinking LLMs in the MT error detection task, measuring their evaluation performance with and without access to the full document context, in the presence and absence of reference translations. Finally, we also explore the effects of fine-tuning across all evaluation settings. We release the code to reproduce our results at `anonymized.com`.

2 Related Work

2.1 Document-level MT evaluation

Most popular MT auto-evaluators operate at the segment level (Rei et al., 2020, 2022; Perrella et al., 2022; Kocmi and Federmann, 2023b,a; Juraska et al., 2023, 2024; Guerreiro et al., 2024).

Recent research has explored incorporating longer contexts in MT evaluation (Vernikos et al., 2022; Raunak et al., 2024; Wang et al., 2025; Domhan and Zhu, 2025). Specifically, Vernikos et al. (2022); Raunak et al. (2024); Wang et al. (2025) design automatic procedures to process longer documents using already existing segment-level auto-evaluators. However, none of these works investigate document-level evaluation by design. They adapt prior segment-level approaches, which are typically based on encoder-only architectures with limited context windows – XLM-RoBERTa, for example, is restricted to 512 tokens (Conneau et al., 2020) – and are therefore unsuited to processing documents beyond relatively short segments. Domhan and Zhu (2025), instead, explore document-level MT evaluation using LLM-as-a-judge. They show that LLM-based MT evaluation is not length-invariant, as auto-evaluators detect fewer error spans in correspondence of longer input texts. However, they do not systematically investigate the performance gains (if any) achieved by document-level evaluation, as opposed to segment-level, nor its relationship with the presence and absence of manually-curated references.

2.2 Reference-free MT evaluation

Neural MT auto-evaluators enable conducting MT evaluation in settings where reference translations are unavailable. Early approaches consist of pre-trained transformer-based models fine-tuned for reference-free evaluation, such as COMET-QE (Rei et al., 2021), CometKiwi (Rei et al., 2022), and MetricX reference-free models (Juraska et al., 2023, 2024, 2025). These models all focus on score-level MT evaluation, returning scalar scores as quality assessments. MaTESe-QE (Perrella et al., 2022) and xCOMET-QE (Guerreiro et al., 2024), instead, are fine-tuned for reference-free MT error detection.

More recently, LLM-as-a-judge approaches have shown their effectiveness in MT evaluation: Kocmi and Federmann (2023a) introduced GEMBA-MQM, an GPT-4-based (OpenAI et al., 2024) auto-evaluator designed to detect translation errors following the Multidimensional Quality Metrics (MQM, Lommel et al., 2014; Freitag et al., 2021) guidelines; similarly, Hrabal et al. (2025) fine-tuned GPT-4o-mini for MQM-based MT evaluation. Yeom et al. (2025) introduced Tagged Span Annotation, an annotation scheme designed to better extract span-level annotations from LLM responses. Juraska et al. (2025) fine-tuned Gemma3 (Team et al., 2025) for span-level MT evaluation using MQM annotations.

Reference-free evaluation enables many additional use cases for MT auto-evaluators. Despite this, reference-free metrics still typically underperform their reference-based counterparts (Freitag et al., 2022b, 2023, 2024; Lavie et al., 2025), suggesting that references contain information that cannot be fully inferred from the source sentence alone. It remains unclear, however, whether document-level context can compensate for the absence of references.

3 Methodology

Two popular human evaluation protocols are Error Span Annotation (ESA, Kocmi et al., 2024b) and MQM. In both protocols, human annotators are presented consecutive segments from the same document and are asked to evaluate them in order. As a result, they produce segment-level evaluations, with access to document-level context.

Our work explores document context LLM-based MT evaluation following the same principle: when incorporating document-level information, we still prompt LLMs to evaluate one segment at a time, but we explicitly instruct them to pay attention to the broader document context provided as input. This design is also motivated by Domhan and Zhu (2025), who recommend the Focus Sentence Prompting (FSP) strategy to mitigate the tendency

	w/ doc context	w/ human references
SL-RF	×	×
SL-RB	×	✓
DL-RF	✓	×
DL-RB	✓	✓

Table 1: Different evaluation scenarios explored in this work. "SL" and "DL" stand for "segment-level" and "document-level", respectively, while "RF" and "RB" stand for "reference-free" and "reference-based", respectively.

of LLMs to detect fewer errors in longer inputs. In FSP, they present the LLM with the full source and translation documents, but prompt it to evaluate a single sentence at a time, rather than evaluating the translation of the whole document in a single pass.[2]

To investigate the impact of document context, also in relation with the presence or absence of reference translations, we design four evaluation scenarios, which we summarize in Table 1. The SL-RF scenario serves as a baseline for our method, where the translation segment is evaluated in isolation by comparing it against the corresponding source segment. In SL-RB, the model is additionally provided with the human reference translation. The DL-RF scenario isolates the impact of document-level context in the absence of human references, providing both the source and translation documents as context, in addition to the source and translation segments. Finally, DL-RB extends DL-RF by also including the human reference translation. This allows us to investigate whether human reference translations and document context provide complementary sources of information.

3.1 Error Detection

Approaches to automatic MT evaluation broadly fall into two categories, depending on their granularity: score-level and span-level. Score-level evaluators assign a scalar quality score, which offers limited interpretability (Perrella et al., 2024), as it conveys little about the nature of the translation errors. Instead, span-level MT evaluation – also referred to as error detection – addresses this limitation by explicitly identifying the spans that contain translation errors and, optionally, assigning to them an error category and severity. For this reason, we conduct our experiments at the span level, instructing LLMs to precisely locate translation errors and classify them according to their severity.

3.1.1 Meta-Evaluation

We measure performance using the MPP (Match with Partial overlap and Partial credit) measure recommended by Perrella et al. (2026). Given a pair of predicted and ground-truth spans, MPP computes precision and recall based on their proportion of overlapping characters. Later, per-span precision and recall are aggregated into their corpus-level counterparts used to compute the final F-score.

Severity Weighting As it stands, MPP measures error detection performance irrespective of error severity. To incorporate severity information into the evaluation, previous works have typically relied on severity penalties (Lavie et al., 2025), which discount the precision and recall obtained in correspondence of errors with mismatched severity. However, as noted by Lyu et al. (2025) and Perrella et al. (2026), this strategy only penalizes severity mismatches, but does not take into account that predicting a minor error where there are no errors is less severe than predicting a major error in the same location. We address this issue by using a severity-based error weighting scheme, rather than a severity penalty that only applies to severity mismatches. Specifically, we re-weight the contribution of `Minor` and `Major` errors to precision and recall using weights 0.2 and 1, reflecting their relative weight of $\frac{1}{5}$ in the MQM protocol.[3] Formally, let us use the same notation as Perrella et al. (2026): predicted and ground-truth error spans are indicated by $\hat{s}$ and s, respectively, whereas $\hat{S}$ and S are sets of predicted and ground-truth error spans across an entire corpus. $M \subseteq \hat{S} \times S$ is a one-to-one matching between predicted and ground-truth error spans. Given a pair $(\hat{s}, s) \in M$, per-span precision and recall – based on proportion of overlapping characters – are $P_{\text{MPP}}(\hat{s}, s)$ and $R_{\text{MPP}}(\hat{s}, s)$. MPP derives corpus-level precision and recall as follows:

$$P_{\text{MPP}} = \frac{1}{|\hat{S}|} \sum_{(\hat{s},s)\in M} P_{\text{MPP}}(\hat{s}, s)$$

$$R_{\text{MPP}} = \frac{1}{|S|} \sum_{(\hat{s},s)\in M} R_{\text{MPP}}(\hat{s}, s)$$

[2] Differently from them, though, we prompt the LLM to evaluate a single *segment* at a time, rather than a single *sentence*. This design choice serves to align our automatic evaluations with the segment-level evaluations available in the test sets employed in this work (Section 4.1).

[3] In MQM, `Minor` and `Major` errors are assigned a penalty of -1 and -5, respectively.

Different from standard MPP, our severity-weighted version rescales per-span precision and recall based on the severity of the involved error spans, before aggregating them into corpus-level statistics:

$$P'_{\text{MPP}}(\hat{s}, s) = P_{\text{MPP}}(\hat{s}, s) \cdot \min(w(\hat{s}), w(s))$$
$$R'_{\text{MPP}}(\hat{s}, s) = R_{\text{MPP}}(\hat{s}, s) \cdot \min(w(\hat{s}), w(s))$$

where $w(\hat{s})$ and $w(s)$ are the weights assigned to $\hat{s}$ and s based on their severity. We compute the final F-score from P'_{MPP} and R'_{MPP}, defined as follows:

$$P'_{\text{MPP}} = \frac{1}{\sum_{\hat{s} \in \hat{S}} w(\hat{s})} \sum_{(\hat{s}, s) \in M} P'_{\text{MPP}}(\hat{s}, s)$$
$$R'_{\text{MPP}} = \frac{1}{\sum_{s \in S} w(s)} \sum_{(\hat{s}, s) \in M} R'_{\text{MPP}}(\hat{s}, s)$$

In this way, correctly predicting a `Minor` error contributes to precision and recall less than correctly predicting a `Major` one. Similarly, predicting a `Minor` error where there are no ground-truth errors is less severe than predicting a `Major` one in the same location. Finally, similar to the severity penalties employed in previous works, severity mismatches between overlapping errors are penalized, because per-span precision and recall are re-scaled using the minimum between the involved spans' severity.

4 Experimental Setup

4.1 Data

We conduct our experiments on the test sets released at WMT24 (Kocmi et al., 2024a; Freitag et al., 2024) and WMT25 (Kocmi et al., 2025; Lavie et al., 2025). These datasets contain annotations of translation quality produced by annotators following the MQM or ESA guidelines. We restrict our experiments to the MQM splits of these two test sets, as the MQM annotators employed at WMT campaigns have been shown to identify more errors than ESA annotators (Kocmi et al., 2024b), making MQM data more suitable for error detection evaluation. Furthermore, we found many error-free ESA translations assigned with final quality scores lower than 100 (perfect score), suggesting that ESA data might be noisier if used for span-based MT evaluation, as opposed to score-based evaluation.

We report statistics about the WMT test sets employed in this work in Table 2.

Test set	LP	#docs	#segs	#spans	$\frac{\#\text{spans}}{\#\text{segs}}$
WMT24	EN→DE	59	6750	8172	1.21
	EN→ES	39	7210	2796	0.39
	JA→ZH	60	5922	4476	0.76
WMT25	EN→KO	24	2574	6500	2.53
	JA→ZH	38	3258	6008	1.84

Table 2: Number of evaluated documents, segments, error spans, and average number of errors per segment in the WMT test sets employed in our experiments.

4.2 Models

Our auto-evaluators consist of three instruction-tuned LLMs with different model scales and context-window capacities: Qwen3-8B, Qwen3-32B (Yang et al., 2025), and Gemma-4-E4B (Team et al., 2024), with context windows of 32K, 32K and 128K, respectively. Through this model selection, our experiments cover two model families, across three different model sizes. Furthermore, all models support a hybrid thinking strategy, meaning that they can operate both in the standard non-thinking scenario and with thinking enabled. This way, we can directly measure the impact of thinking on MT error detection across our evaluation scenarios.

4.3 Fine-tuning

To the best of our knowledge, all previous approaches fine-tuning models for MT evaluation operate at the segment level. This work addresses this gap by also exploring the impact of incorporating the broader document context into fine-tuning. Specifically, we fine-tune the Qwen3-8B model under all four evaluation scenarios in Table 1. Details about our fine-tuning experimental setup are available in Appendix D.

5 Results

Table 3 reports the MPP F-score on the WMT24 and WMT25 MQM datasets across different models and evaluation scenarios. As recommended by Perrella et al. (2026), we only report MPP results micro-averaged across test set samples, since macro-averaging results was found to be biased to prefer auto-evaluators that detect fewer error spans. A breakdown of per-language pair results is provided in Tables 8 and 9 in Appendix C.

We contextualize the results of our auto-evaluators in the broader span-level evaluation landscape by also reporting the performance achieved

Test set	Model	Think	Evaluation Scenarios			
			SL-RF	SL-RB	DL-RF	DL-RB
WMT24 EN→DE EN→ES JA→ZH	*Baseline*					
	xCOMET-XL	×	9.58	11.78	–	–
	Qwen3					
	Qwen3-8B	×	15.46	16.09	15.98	16.34
	Qwen3-8B	✓	20.19	18.95	21.73	20.25
	Qwen3-32B	×	20.60	19.99	22.41	20.81
	Qwen3-32B	✓	23.21	20.87	**24.54**	21.86
	Gemma-4					
	Gemma-4-E4B	×	21.65	21.44	22.17	21.74
	Gemma-4-E4B	✓	20.85	20.32	22.52	22.58
	Qwen3 Finetuned					
	Qwen3-8B-finetuned	×	16.75	19.68	20.16	21.69
WMT25 EN→KO JA→ZH	*Baseline*					
	xCOMET-XL	×	12.93	15.40	–	–
	Qwen3					
	Qwen3-8B	×	14.79	18.96	12.12	14.97
	Qwen3-8B	✓	14.69	19.63	10.95	14.70
	Qwen3-32B	×	20.43	**23.00**	18.28	21.01
	Qwen3-32B	✓	13.81	20.94	12.49	19.13
	Gemma-4					
	Gemma-4-E4B	×	17.02	21.01	11.28	12.97
	Gemma-4-E4B	✓	13.15	18.09	11.21	13.77
	Qwen3 Finetuned					
	Qwen3-8B-finetuned	×	17.36	19.12	16.47	18.74

Table 3: Weighted micro-averaged MPP F-score on the WMT24 and WMT25 MQM test sets across different models and evaluation scenarios. The "Think" column indicates whether the model was run with thinking enabled. The best result per test set is in bold, while the best result per model is underlined.

by xCOMET-XL (Guerreiro et al., 2024). Our models outperform this baseline in nearly all evaluation settings.

As shown in Table 3, incorporating document context into the evaluation has mixed effects that depend on the test set. In particular, it improves performance across the board on WMT24, while appearing to deteriorate performance on WMT25. On WMT24, the highest scores are largely achieved by DL-RF, which may suggest that document-level information in this test set is sufficient to eliminate the need for reference translations. However, the same document-level, reference-free evaluation scenario yields the lowest performance on WMT25, where SL-RB achieves the best results.

These results call for a deeper investigation into the impact of document-level information on evaluation: Is document-level information unnecessary or irrelevant for the MT evaluation task on this test set? Are there other confounding factors contributing to the observed performance decrease? More generally, why do our results differ so substantially between the WMT24 and WMT25 test sets?

To address these questions, we first examine the main structural differences between the two test sets. We then analyze how performance varies across three additional dimensions: the availability of human reference translations, the role of thinking, and task-specific fine-tuning (Figure 1).

5.1 Structural Differences Between WMT24 and WMT25

In both the WMT24 and WMT25 test sets, segments are multi-sentence units of text obtained by splitting larger documents (Kocmi et al., 2024a, 2025). Despite sharing the same definition, both segments and documents in WMT24 are substantially shorter than those in WMT25. To illustrate this difference, we report character-length statistics for the two test sets in Table 5.

The difference in length is particularly pronounced in EN→KO, where English source documents span 18,415.9 characters on average, which is almost an order of magnitude longer than the average English source documents in WMT24 EN→ES and EN→DE. Similarly, documents in

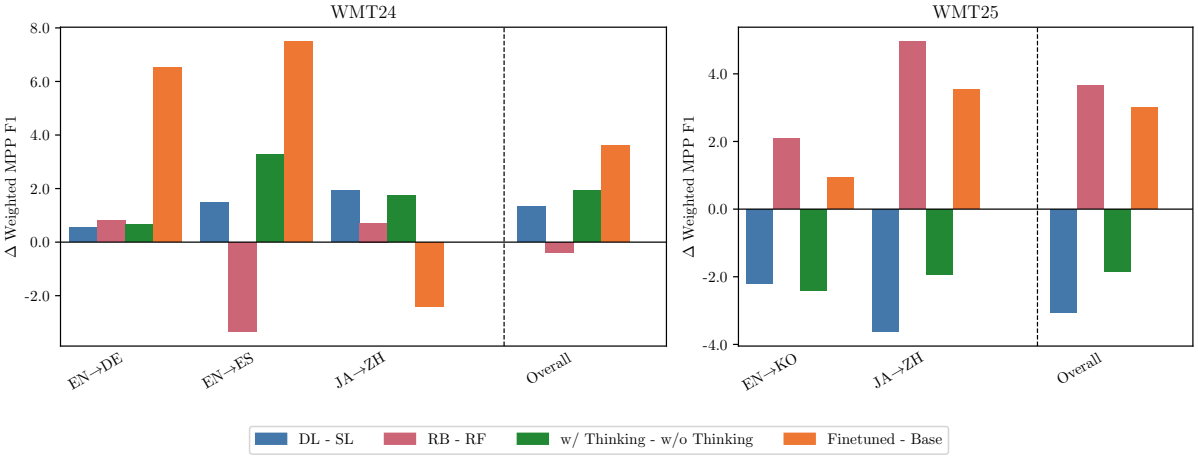


Figure 1: Average marginal effects on weighted MPP F-score across language pairs and test sets. Bars indicate the F-score differences associated with document-level versus segment-level evaluation (DL-SL), reference-based versus reference-free (RB-RF), enabling thinking versus disabling it, and after fine-tuning relative to the base model.

WMT25 JA→ZH are, on average, four times longer than those in WMT24 JA→ZH. Comparable differences, although less pronounced, can also be observed when comparing the lengths of individual segments.

Another important distinction between the two test sets is the difficulty-based subsampling strategy used to construct the WMT25 test set (Kocmi et al., 2025). Specifically, the WMT25 test set was designed to include documents considered difficult for automatic systems to translate. To achieve this, the WMT organizers adopted the sentinel-src difficulty estimator introduced by Proietti et al. (2025). In contrast, documents in the WMT24 test set were randomly sampled. As a result, the two test sets differ not only in terms of segment and document lengths, but also substantially in composition, since WMT25 contains texts that were explicitly selected to maximize translation difficulty.

5.2 Impact of Human References

Figure 1 shows the marginal effect of document context, reference translations, thinking, and fine-tuning on evaluation performance. As we can see, manually curated references provide little benefit on WMT24 and even reduce performance in EN→ES. By contrast, results on WMT25 show consistent performance improvements when reference translations are available. We hypothesize that this difference stems from the interaction between task complexity, reference bias, and source text difficulty.

RF evaluation is more complex as it requires auto-evaluators to compare the translation with the source text in a different language. Including reference translations, however, has a mixed effect: while it simplifies the evaluation task – reducing it to a same-language comparison between texts – it also introduces evaluation bias, whereby translations more similar to the references may be preferred. This issue is particularly problematic when references are of low quality.

Our hypothesis is that the evaluated models are better able to handle the RF evaluation task in WMT24, where source texts are easier (and shorter). As a result, performance remains roughly unchanged and sometimes even decreases when including references, as observed for EN→ES. This particular result is likely due to the lower quality of the Spanish reference translations, as reported by Freitag et al. (2024), who show that the Spanish reference was judged to be of lower quality than the translations produced by several automatic systems (Table 5 of the findings paper). In contrast, the references for EN→DE and JA→ZH were both ranked among the top systems (i.e., within the first cluster of statistical significance).

Conversely, with increased source text difficulty, evaluation performance on WMT25 benefits substantially from the inclusion of reference translations, which reduce overall task complexity. This interpretation aligns with findings from the WMT25 Shared Task on Automated Translation Evaluation Systems, where reference translations proved to be particularly beneficial, compared to previous years, across all tested auto-evaluators (Lavie et al., 2025).

These considerations also partly extend to the

impact of document context. Although document-level information appears to be helpful for the MT evaluation task on a simpler test set such as WMT24, it may overall increase task complexity by requiring LLMs to process longer inputs, which are often sparse in terms of information actually relevant to the evaluation task. This could explain why our results differ so markedly between the WMT24 and WMT25 test sets.

5.3 Impact of Thinking and the Number of Predicted Spans

Similar to reference translations, Figure 1 shows that thinking improves performance on WMT24, while it consistently reduces performance on WMT25. In this respect, and also contributing to explain the impact of reference translations and document context, another factor that we find to affect the performance differences between WMT24 and WMT25 is the impact that different evaluation scenarios have on the number of error spans produced by auto-evaluators. We investigate this issue in Figure 4 in Appendix E.

A consistent pattern emerges across both test sets and models: both document-level information and the thinking inference mode lead models to detect fewer errors than the corresponding segment-level and non-thinking approaches. In contrast, the presence of reference translations increases the number of detected errors. This provides another explanation for why document-level information and thinking deteriorate performance on WMT25. Indeed, as shown in Table 2, WMT25 contains substantially more errors per segment (likely due to its higher difficulty). As a result, variations in the experimental setup whose byproduct is an increase (or decrease) in the number of detected error spans may appear to improve (or deteriorate) performance depending on the specific characteristics of the test set under consideration. In our context, document-level evaluation and thinking may therefore appear to disproportionately deteriorate performance on the WMT25 test set.

This effect is also reflected in the precision–recall trade-off achieved by different evaluation scenarios, which we report in Tables 6 and 7 in Appendix C. Notably, DL-RF and DL-RB, especially when thinking is enabled, exhibit particularly high precision values, but their very low recall drags the overall F-score down on WMT25.

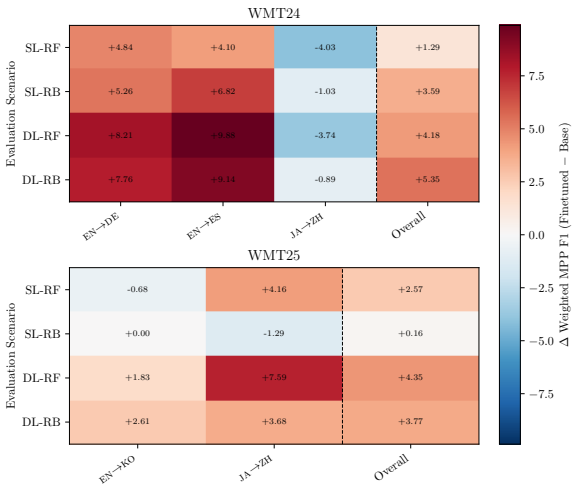


Figure 2: Averaged impact of fine-tuning on weighted MPP F-score across evaluation scenarios, datasets, and language pairs. Positive values indicate gains from fine-tuning, while negative values indicate drops.

5.4 Fine-Tuning for Document-Level Evaluation

We find task-specific fine-tuning to improve performance across most evaluation scenarios and test sets. As shown in Figure 2, the gains are especially evident when document-level context is available: improvements over the base model are generally larger in DL-RF and DL-RB than in the corresponding segment-level settings.

This finding is particularly relevant in light of the results on the WMT25 test set. As discussed in previous sections, document-level context has a notably negative impact on WMT25 performance, which we attributed primarily to the length of WMT25 documents, the difficulty of their source texts, and the effect of document-level evaluation on the number of error spans that auto-evaluators detect. Nonetheless, fine-tuning for document-level MT evaluation yields clear improvements on this test set as well. Furthermore, fine-tuning improves performance more when document context is available (DL-RF and DL-RB), compared to segment-level fine-tuning (SL-RF and SL-RB). This finding allows us to rule out the hypothesis that document-level context is uninformative for WMT25. The fact that our fine-tuned model learned to better leverage document-level information suggests that the additional context is indeed relevant to the task, but that models may struggle to exploit it effectively when dealing with more challenging test sets and longer source texts.

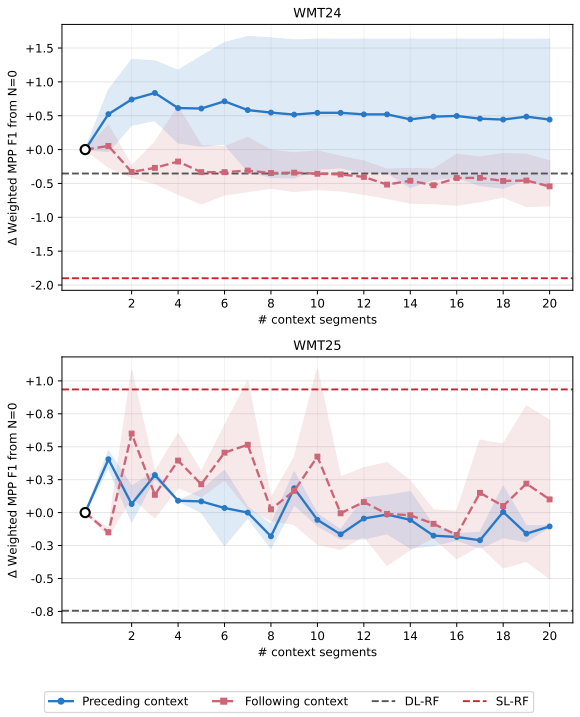


Figure 3: Average effect of adding preceding and following context paragraphs on MPP F-score, relative to no context (N=0). Each curve is the mean over WMT24 language pairs and WMT25 language pairs; shaded bands show the per-language-pair range.

6 Is the Whole Document Needed?

In this section, we investigate whether full-document context is actually necessary, or whether chunking the input document into smaller pieces can provide the relevant information while reducing task complexity.

We explore this by measuring the MPP F-score while varying the number of segments provided as evaluation context to Qwen3-32B. We distinguish between preceding and following context, and report results Figure 3, while a per-language breakdown is provided in Appendix G. All evaluations are conducted without reference translations and with thinking disabled. To contextualize these results with respect to our previous findings, we also report the performance of DL-RF, i.e., reference-free evaluation conducted with the whole document as input, and SL-RF, i.e., reference-free evaluation conducted at the segment level.

On WMT24, evaluation performance increases when preceding segments are provided as additional context in the input. However, the gains saturate quickly, at roughly three preceding segments, beyond which additional context leaves performance unchanged or slightly degrades it. By contrast, performance consistently decreases when the additional segments follow the segment being evaluated.

On WMT25, however, variance appears to dominate the curves, and the effects of preceding and following segments remain unclear. We hypothesize that this stems from the considerations discussed above: the model may not be able to effectively leverage the additional context in the WMT25 test set, making it difficult to measure the more fine-grained effect of individual additional segments on evaluation performance. Indeed, SL-RF achieves the highest performance on WMT25, whereas DL-RF yields the lowest performance.

7 Conclusion

This work systematically examines how document context affects LLM-based MT evaluation and how this interacts with manually curated references. On WMT24, document-level evaluation consistently outperforms segment-level evaluation, while references rarely help and often hurt performance, regardless of evaluation granularity. In contrast, on WMT25, document-level evaluation consistently underperforms segment-level evaluation, and references always improve results.

To explain these contrasting findings, we analyze test-set features that may influence evaluation outcomes. We hypothesize that longer and more difficult source documents make evaluation substantially more complex, especially when full document context is included. In such settings, references simplify the task and improve performance. We also find that document-level evaluation and thinking inference tend to reduce the number of detected errors, which may hurt performance on WMT25, where segments contain more errors on average.

Finally, fine-tuning improves results on both WMT24 and WMT25, especially when performed at the document level. This suggests that document-level evaluation may be superior in principle, but that the tested base models struggle to exploit document-level information on harder benchmarks such as WMT25. Future work should test whether document-level fine-tuning scales to larger models, potentially enabling state-of-the-art MT evaluation through better use of document context.

Limitations

Selection of Models Our study covers three LLMs across two model families, spanning three distinct model sizes. We selected these models because they are widely used and both support a hybrid thinking strategy, allowing us to compare thinking and non-thinking performance using the same models. Due to budget constraints, we were unable to experiment with larger open-source models or closed-source alternatives.

Selection of Translation Directions Our experiments cover only four translation directions: English to German, English to Korean, English to Spanish, and Japanese to Chinese. This selection is limited by the availability of MQM annotations in the WMT24 and WMT25 data. We restricted our analysis to these two test sets, excluding older available MQM data, to avoid running our experiments on data that had a high chance of being included in the LLMs' pre-training data.

Acknowledgments

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, and 17 others. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Duarte Miguel Alves, José Pombal, Nuno M Guerreiro, Pedro Henrique Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and Andre Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. In *First Conference on Language Modeling*.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, and 2 others. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Tobias Domhan and Dawei Zhu. 2025. Same evaluation, more tokens: On the effect of input length for machine translation evaluation using large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7940–7958, Suzhou, China. Association for Computational Linguistics.

Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and Andre Martins. 2022. Quality-aware decoding for neural machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1396–1412, Seattle, United States. Association for Computational Linguistics.

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.

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2022a. High quality rather than high model probability: Minimum bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio,

Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022b. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Nuno M. Guerreiro, 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.

Miroslav Hrabal, Ondrej Glembek, Aleš Tamchyna, Almut Silja Hildebrand, Alan Eckhard, Miroslav Štola, Sergio Penkale, Zuzana Šimečková, Ondřej Bojar, Alon Lavie, and Craig Stewart. 2025. CUNI and phrase at WMT25 MT evaluation task. In *Proceedings of the Tenth Conference on Machine Translation*, pages 934–944, Suzhou, China. Association for Computational Linguistics.

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.

Juraj Juraska, 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 *Proceedings of the Tenth Conference on Machine Translation*, pages 957–968, Suzhou, China. Association for Computational Linguistics.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Christof Monz, Kenton Murray, and 10 others. 2025. Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413, Suzhou, China. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, and 3 others. 2024a. 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, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, and 3 others. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023a. 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.

Tom Kocmi and Christian Federmann. 2023b. 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, Tampere, Finland. European Association for Machine Translation.

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

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Alon Lavie, 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 *Proceedings of the Tenth Conference on Machine Translation*, pages 436–483, Suzhou, China. Association for Computational Linguistics.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. 2025. Prompt repetition improves non-reasoning llms. *Preprint*, arXiv:2512.14982.

Arle Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica: tecnologies de la traducció*, 0:455–463.

Boxuan Lyu, Haiyue Song, Hidetaka Kamigaito, Chenchen Ding, Hideki Tanaka, Masao Utiyama, Kotaro Funakoshi, and Manabu Okumura. 2025. Minimum bayes risk decoding for error span detection in reference-free automatic machine translation evaluation. *Preprint*, arXiv:2512.07540.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Stefano Perrella, Eric Morales Agostinho, and Hugo Zaragoza. 2026. Span-level machine translation meta-evaluation. *Preprint*, arXiv:2603.19921.

Stefano Perrella, Lorenzo Proietti, Pere-Lluís Huguet Cabot, Edoardo Barba, and Roberto Navigli. 2024. Beyond correlation: Interpretable evaluation of machine translation metrics. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20689–20714, Miami, Florida, USA. Association for Computational Linguistics.

Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Niccolò Campolungo, and Roberto Navigli. 2022. MaTESe: Machine translation evaluation as a sequence tagging problem. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 569–577, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Jan-Thorsten Peter, David Vilar, Daniel Deutsch, Mara Finkelstein, Juraj Juraska, and Markus Freitag. 2023. There's no data like better data: Using QE metrics for MT data filtering. In *Proceedings of the Eighth Conference on Machine Translation*, pages 561–577, Singapore. Association for Computational Linguistics.

Lorenzo Proietti, Stefano Perrella, Vilém Zouhar, Roberto Navigli, and Tom Kocmi. 2025. Estimating machine translation difficulty. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 24261–24285, Suzhou, China. Association for Computational Linguistics.

Miguel Ramos, Patrick Fernandes, António Farinhas, and Andre Martins. 2024. Aligning neural machine translation models: Human feedback in training and inference. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 258–274, Sheffield, UK. European Association for Machine Translation (EAMT).

Vikas Raunak, Tom Kocmi, and Matt Post. 2024. SLIDE: Reference-free evaluation for machine translation using a sliding document window. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 205–211, Mexico City, Mexico. Association for Computational Linguistics.

Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. Are references really needed? unbabel-IST 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online. 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.

Ricardo Rei, 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 *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Gemma Team, 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, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, and 89 others. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Giorgos Vernikos, Brian Thompson, Prashant Mathur, and Marcello Federico. 2022. Embarrassingly easy document-level MT metrics: How to convert any pretrained metric into a document-level metric. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 118–128, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Kuang-Da Wang, Shuoyang Ding, Chao-Han Huck Yang, Ping-Chun Hsieh, Wen-Chih Peng, Vitaly Lavrukhin, and Boris Ginsburg. 2025. Extending automatic machine translation evaluation to book-length documents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32311–32327, Suzhou, China. Association for Computational Linguistics.

Billy T. M. Wong and Chunyu Kit. 2012. Extending machine translation evaluation metrics with lexical cohesion to document level. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1060–1068, Jeju Island, Korea. Association for Computational Linguistics.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: pushing the boundaries of llm performance in machine translation. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Taemin Yeom, Yonghyun Ryu, Yoonjung Choi, and Jinyeong Bak. 2025. Tagged span annotation for detecting translation errors in reasoning LLMs. In *Proceedings of the Tenth Conference on Machine Translation*, pages 878–886, Suzhou, China. Association for Computational Linguistics.

Mao Zheng, Zheng Li, Yang Du, Bingxin Qu, and Mingyang Song. 2025. Shy-hunyuan-MT at WMT25 general machine translation shared task. In *Proceedings of the Tenth Conference on Machine Translation*, pages 607–613, Suzhou, China. Association for Computational Linguistics.

A Prompt Templates

System prompt

```
You are an expert annotator tasked with evaluating the quality of machine translation.
```

User prompt

```
{source_lang} Source Document:
```{source_document}```

{source_lang} Source Segment:
```{source_segment}```

{target_lang} Translation Document:
```{translation_document}```

{target_lang} Translation Segment:
```{translation_segment}```

{target_lang} Reference Segment:
```{reference_segment}```

You are given a source segment and its machine translation segment, enclosed in triple backticks. Your task is to identify and classify the errors in the translation segment by comparing it to the source segment.

You are given a source segment, its human reference segment, and its machine translation segment, enclosed in triple backticks. Your task is to identify and classify the errors in the translation segment by comparing it to the source segment and the reference segment.

You are given a source segment, its machine translation segment, and their corresponding full documents, enclosed in triple backticks. Your task is to identify and classify the errors in the translation segment by comparing it to the source segment, while using the full documents to resolve contextual ambiguities.

You are given a source segment, its human reference segment, its machine translation segment, and the corresponding full source and translation documents, enclosed in triple backticks. Your task is to identify and classify the errors in the translation segment by comparing it to the source segment, while using the full documents to resolve contextual ambiguities.

The categories and subcategories of errors are: accuracy (addition, mistranslation, omission, untranslated text), fluency (character encoding, grammar, inconsistency, punctuation, register, spelling), locale convention (currency, date, name, telephone, or time format), style (awkward), terminology (inappropriate for context, inconsistent use), non-translation, or other.

Each error is classified by severity:
- critical: completely changes the meaning or inhibits comprehension of the text.
- major: notably changes the meaning or disrupts readability.
- minor: slightly changes the meaning or affects readability.

For each error, provide:
1. The minimal error span: the smallest contiguous string in the translation that contains the error, without including surrounding correct text.
2. The extended error span: the minimal error span extended by a few words before and after, taken directly from the translation as they appear.
3. The error type (error category and subcategory).
4. The error severity (critical, major or minor).

Use this exact format in your final answer:

{
  "errors": [
    {
      "minimal_span": "<minimal span of text that contains the error>",
      "extended_span": "<extended span of text that contains the error>",
      "type": "<error category>/<error subcategory>",
      "severity": "<critical|major|minor>"
    }
  ]
}

If no errors are found, return {"errors": []}.
```

Table 4: Unified prompt template used across all evaluation scenarios. Gray-shaded fields are included only when required by the selected scenario: reference segments in RB settings, and source-translation documents in DL settings. Colored instruction blocks correspond to the scenario-specific instructions for SL-RF, SL-RB, DL-RF, and DL-RB. All other prompt components, including the output format, error taxonomy, and severity definitions, are shared across scenarios.
```

B Test Set Statistics

Table 5 provides character-level statistics on segment and document lengths for the WMT24 and WMT25 test sets. These statistics help contextualize the discussion in Section 5.1, showing that the two benchmarks differ substantially not only in terms of language pairs, but also in their structure.

C Additional Results

This appendix provides additional results that complement the aggregate MMP F-score results reported in Table 3. Specifically, Tables 6 and 7 report weighted MPP precision and recall scores, respectively. In addition, Tables 8 and 9 present a breakdown by language pair and evaluation scenario of weighted MPP F-score, which enables us to assess whether the trends observed in the main results hold consistently across individual translation directions.

D Fine-tuning Configuration

Table 10 provides the details of our fine-tuning setup. We fine-tune the model separately under each evaluation scenario, using the same input structure and output format adopted at evaluation time.

E Analysis of Predicted Spans

Tables 4 and 5 report analyses of the spans predicted by our automatic evaluators. We examine both the number of predicted error spans and their average character length across models, datasets, and evaluation scenarios.

F Average Impact of Document Context and Thinking

This appendix reports matched comparisons that isolate the impact of individual dimensions. Figure 6 shows the impact of adding document-level context, while Figure 7 shows the impact of adding thinking.

G Context Window Study

Figures 8 and 9 show how performance varies with the size and position of the document context provided to the model for WMT24 and WMT25, respectively.

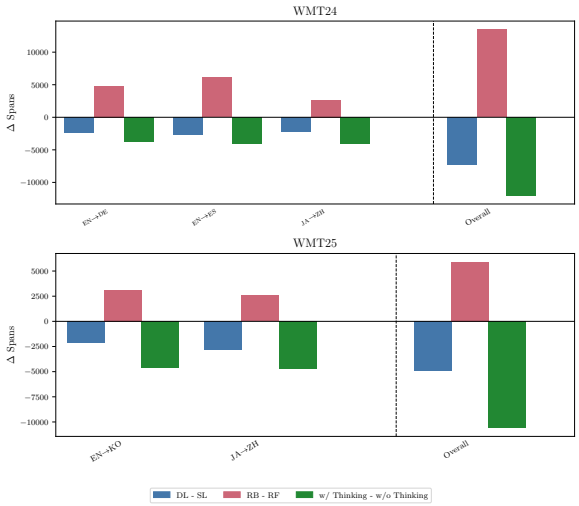


Figure 4: Average matched marginal effects on the total number of predicted spans across language pairs and test sets. Positive values indicate that a setting produces more predicted spans, while negative values indicate fewer predicted spans.

H Prompt Repetition

We highlight an interesting phenomenon that emerges from these results. First, note that `SL-RF` differs from the point located at position 0 on the x-axis. The latter corresponds to a segment-level evaluation prompt, whereas in the former setting we prompt the LLM with a document-level evaluation prompt in which the source document is identical to the source segment (both evaluation prompts are reported in Appendix X). Interestingly, `SL-RF` underperforms the point at index 0 on the WMT24 test set. This may seem unintuitive, since the two evaluation prompts provide exactly the same information to the model. We hypothesize that this is related to the findings of Leviathan et al. (2025), who show that repeating the evaluation prompt twice systematically increases performance across a wide range of tasks. As a result, simply repeating the source segment, both as the source segment and as document context, may improve evaluation performance in segment-level evaluation scenarios. However, this pattern does not also hold for the WMT25 test set. Here again, we hypothesize that this is due to length and source-text difficulty, and that repeating a long source text that is difficult to translate may further complicate the evaluation task.

Test set	LP	Stat.	Segment Length	Document Length
WMT24	EN→DE	avg	154.2	2084.0
		std	158.3	1528.4
		med	91.5	1433.0
		max	886.0	5786.0
	EN→ES	avg	186.4	3051.3
		std	177.7	1791.3
		med	114.5	2650.0
		max	967.0	5667.0
	JA→ZH	avg	87.4	997.8
		std	78.1	483.3
		med	67.5	920.0
		max	531.0	1701.0
WMT25	EN→KO	avg	575.0	18415.9
		std	220.2	11937.6
		med	620.0	28130.0
		max	1057.0	28560.0
	JA→ZH	avg	181.3	4134.3
		std	139.0	4630.3
		med	149.0	802.0
		max	600.0	11272.0

Table 5: Character-length statistics for WMT24 and WMT25 test sets by language pair. Segment Length and Document Length denote source-side segment- and document-level lengths in characters.

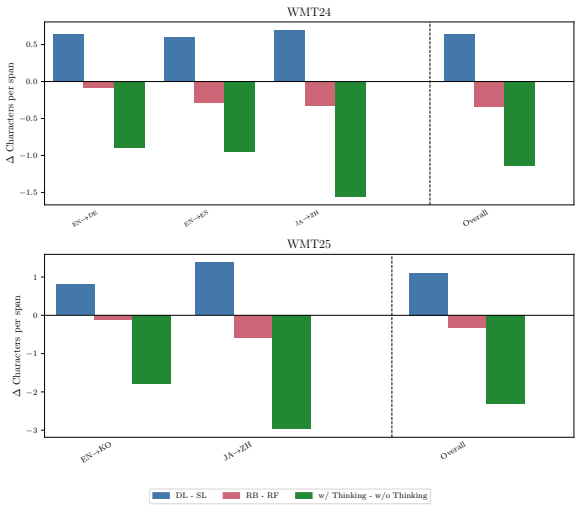


Figure 5: Average matched marginal effects on the average number of characters per predicted span across language pairs and test sets. Positive values indicate longer predicted spans, while negative values indicate shorter predicted spans.

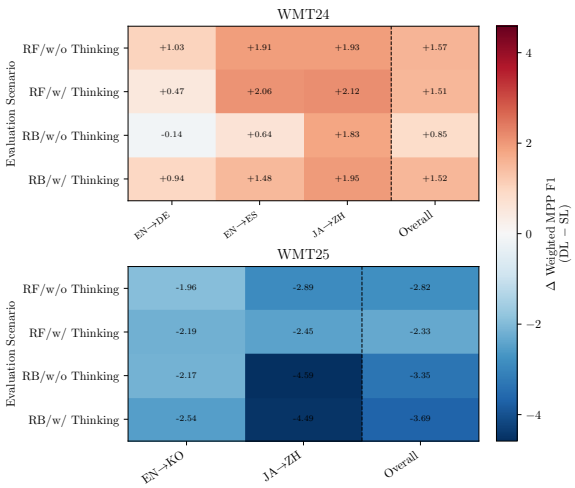


Figure 6: Average impact of document-level prompting over segment-level prompting on weighted MPP F-score across different reference condition, datasets, and language pairs.

Test set	Model	Think	Evaluation Scenarios			
			SL-RF	SL-RB	DL-RF	DL-RB
WMT24 EN→DE EN→ES JA→ZH	*Baseline*					
	xCOMET-XL	×	5.54	6.93	-	-
	Qwen3					
	Qwen3-8B	×	10.14	9.75	10.56	10.32
	Qwen3-8B	✓	17.03	12.92	20.44	15.26
	Qwen3-32B	×	14.08	12.60	16.58	13.85
	Qwen3-32B	✓	20.97	14.18	**25.31**	15.53
	Gemma-4					
	Gemma-4-E4B	×	17.24	14.67	22.43	18.59
	Gemma-4-E4B	✓	19.60	15.17	24.66	19.59
	Qwen3 Finetuned					
	Qwen3-8B-finetuned	×	10.73	13.11	14.02	15.30
WMT25 EN→KO JA→ZH	*Baseline*					
	xCOMET-XL	×	8.78	11.17	-	-
	Qwen3					
	Qwen3-8B	×	12.69	12.94	10.75	12.24
	Qwen3-8B	✓	21.46	20.04	24.46	21.83
	Qwen3-32B	×	18.01	18.80	16.57	18.79
	Qwen3-32B	✓	25.35	22.98	**30.21**	23.78
	Gemma-4					
	Gemma-4-E4B	×	21.70	19.57	26.38	22.68
	Gemma-4-E4B	✓	24.00	22.12	29.68	26.72
	Qwen3 Finetuned					
	Qwen3-8B-finetuned	×	11.79	14.10	13.72	17.03

Table 6: Weighted micro-averaged MPP precision on the WMT24 and WMT25 MQM test sets across different models and evaluation scenarios. The "Think" column indicates whether the model was run with reasoning enabled. The best result per test set is in bold, while the best result per model is underlined.

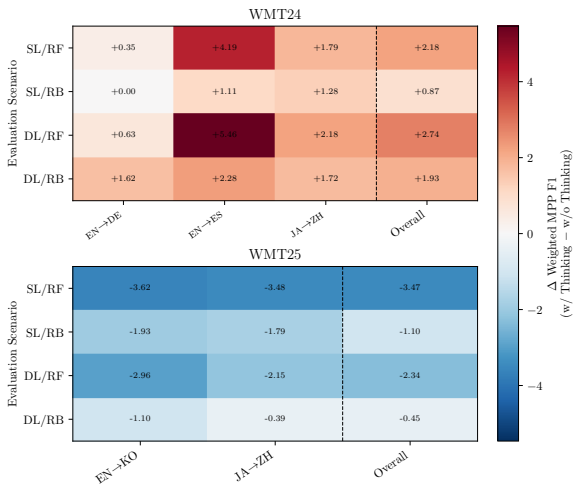


Figure 7: Averaged impact of Thinking over Non-Thinking strategy on weighted MPP F-score across prompting scenarios, datasets, and language pairs. Positive values indicate gains from thinking, while negative values indicate drops.

Test set	Model	Think	Evaluation Scenarios			
			SL-RF	SL-RB	DL-RF	DL-RB
WMT24 EN→DE EN→ES JA→ZH	*Baseline*					
	xCOMET-XL	×	35.47	39.04	-	-
	Qwen3					
	Qwen3-8B	×	32.51	45.98	32.85	39.19
	Qwen3-8B	✓	24.78	35.53	23.20	30.08
	Qwen3-32B	×	38.40	**48.32**	34.57	41.80
	Qwen3-32B	✓	25.97	39.45	23.81	36.91
	Gemma-4					
	Gemma-4-E4B	×	29.09	39.77	21.91	26.17
	Gemma-4-E4B	✓	22.27	30.78	20.72	26.64
	Qwen3 Finetuned					
	Qwen3-8B-finetuned	×	38.13	39.45	35.82	37.24
WMT25 EN→KO JA→ZH	*Baseline*					
	xCOMET-XL	×	24.52	24.78	-	-
	Qwen3					
	Qwen3-8B	×	17.23	29.23	13.87	19.29
	Qwen3-8B	✓	11.16	19.23	7.05	11.09
	Qwen3-32B	×	23.61	**29.64**	20.39	23.81
	Qwen3-32B	✓	9.49	19.24	7.87	16.01
	Gemma-4					
	Gemma-4-E4B	×	14.00	22.68	7.17	9.09
	Gemma-4-E4B	✓	9.06	15.31	6.91	9.27
	Qwen3 Finetuned					
	Qwen3-8B-finetuned	×	32.94	29.70	20.60	20.83

Table 7: Weighted micro-averaged MPP recall on the WMT24 and WMT25 MQM test sets across different models and evaluation scenarios. The "Think" column indicates whether the model was run with reasoning enabled. The best result per test set is in bold, while the best result per model is underlined.

LP	Model	Think	Evaluation Scenarios			
			SL-RF	SL-RB	DL-RF	DL-RB
EN→DE	*Baseline*					
	xCOMET-XL	×	15.63	17.59	-	-
	Qwen3					
	Qwen3-8B	×	17.48	19.20	17.65	18.19
	Qwen3-8B	✓	20.34	21.22	21.08	21.13
	Qwen3-32B	×	24.18	23.82	25.19	24.30
	Qwen3-32B	✓	23.76	23.33	23.63	25.05
	Gemma-4					
	Gemma-4-E4B	×	21.83	23.59	21.21	22.07
	Gemma-4-E4B	✓	20.43	22.06	21.22	23.25
	Qwen3 Finetuned					
	Qwen3-8B-finetuned	×	22.32	24.46	25.86	**25.95**
EN→ES	*Baseline*					
	xCOMET-XL	×	5.62	7.58	-	-
	Qwen3					
	Qwen3-8B	×	9.32	9.08	8.81	8.42
	Qwen3-8B	✓	15.69	11.24	17.00	12.46
	Qwen3-32B	×	15.93	13.13	17.95	13.62
	Qwen3-32B	✓	20.79	13.95	**23.01**	14.23
	Gemma-4					
	Gemma-4-E4B	×	16.89	13.32	17.74	14.39
	Gemma-4-E4B	✓	18.23	13.66	20.88	16.59
	Qwen3 Finetuned					
	Qwen3-8B-finetuned	×	13.42	15.90	18.69	17.56
JA→ZH	*Baseline*					
	xCOMET-XL	×	9.88	12.08	-	-
	Qwen3					
	Qwen3-8B	×	18.25	18.53	20.46	21.17
	Qwen3-8B	✓	22.50	22.46	24.66	25.02
	Qwen3-32B	×	21.02	21.51	22.63	22.29
	Qwen3-32B	✓	23.69	23.56	**25.91**	24.40
	Gemma-4					
	Gemma-4-E4B	×	24.06	25.15	25.44	26.25
	Gemma-4-E4B	✓	22.52	23.02	24.51	25.46
	Qwen3 Finetuned					
	Qwen3-8B-finetuned	×	14.22	17.50	16.72	20.28

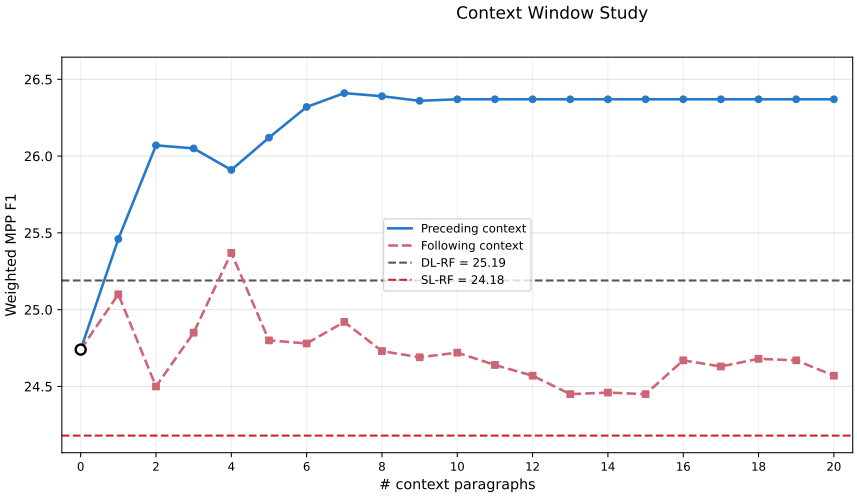
Table 8: Weighted micro-averaged MPP F-score on the WMT24 test set across different models and evaluation scenarios. The "Think" column indicates whether the model was run with reasoning enabled. The best result per test set is in bold, while the best result per model is underlined.

LP	Model	Think	Evaluation Scenarios			
			SL-RF	SL-RB	DL-RF	DL-RB
EN→KO	*Baseline*					
	xCOMET-XL	×	8.76	10.03	-	-
	Qwen3					
	Qwen3-8B	×	14.26	14.24	12.45	13.06
	Qwen3-8B	✓	12.50	14.64	10.52	11.71
	Qwen3-32B	×	17.36	19.31	15.32	17.22
	Qwen3-32B	✓	13.10	16.67	11.24	16.13
	Gemma-4					
	Gemma-4-E4B	×	16.39	**19.39**	11.69	12.54
	Gemma-4-E4B	✓	11.56	15.83	8.83	11.68
	Qwen3 Finetuned					
	Qwen3-8B-finetuned	×	13.58	14.24	14.28	15.67
JA→ZH	*Baseline*					
	xCOMET-XL	×	15.33	18.73	-	-
	Qwen3					
	Qwen3-8B	×	14.66	22.62	11.02	16.97
	Qwen3-8B	✓	15.92	22.75	11.20	16.50
	Qwen3-32B	×	22.64	**26.69**	21.32	23.69
	Qwen3-32B	✓	14.29	23.80	13.26	21.15
	Gemma-4					
	Gemma-4-E4B	×	17.37	22.12	11.00	13.11
	Gemma-4-E4B	✓	14.03	19.51	12.44	14.94
	Qwen3 Finetuned					
	Qwen3-8B-finetuned	×	18.82	21.33	18.61	20.65

Table 9: Weighted micro-averaged MPP F-score on the WMT25 test set across different models and evaluation scenarios. The "Think" column indicates whether the model was run with reasoning enabled. The best result per test set is in bold, while the best result per model is underlined.

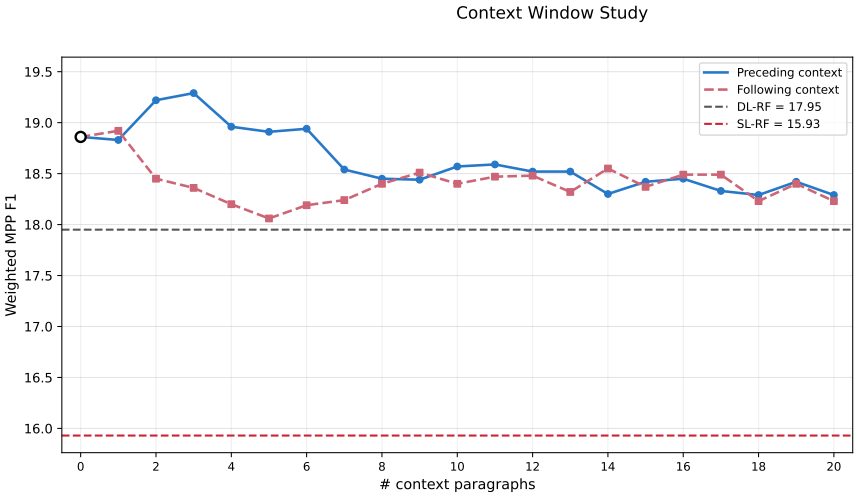
Training configuration	
Learning rate	1×10^{-4}
Weight decay	0.01
Batch size	8
Epoch	1
Optimizer	AdamW
LR scheduler	Linear
Test split	0.1
LoRA configuration	
Rank (r)	16
Alpha (α)	32
Dropout	0.05
Training data	
WMT22	EN→DE, ZH→EN, EN→ZH, EN→RU
WMT23	EN→DE, ZH→EN
Total	89,480 samples

Table 10: Fine-tuning configuration used in our experiments.



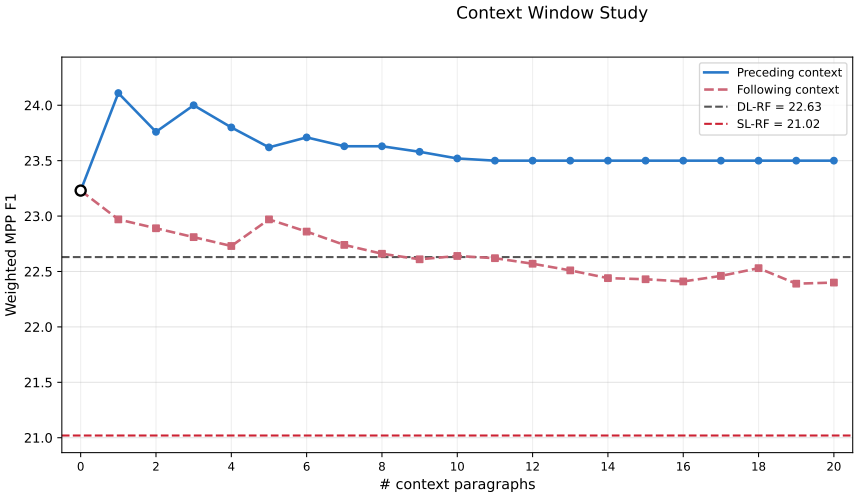
N	DL-src (P)	DL-src (F)
0	154	154
2	364	492
4	463	781
6	505	994
8	518	1143
10	522	1243
12	522	1315
14	522	1371
16	522	1415
18	522	1454
20	522	1493

(a) EN→DE



N	DL-src (P)	DL-src (F)
0	186	186
2	509	531
4	756	796
6	935	993
8	1063	1134
10	1154	1235
12	1220	1308
14	1274	1367
16	1322	1418
18	1363	1464
20	1399	1503

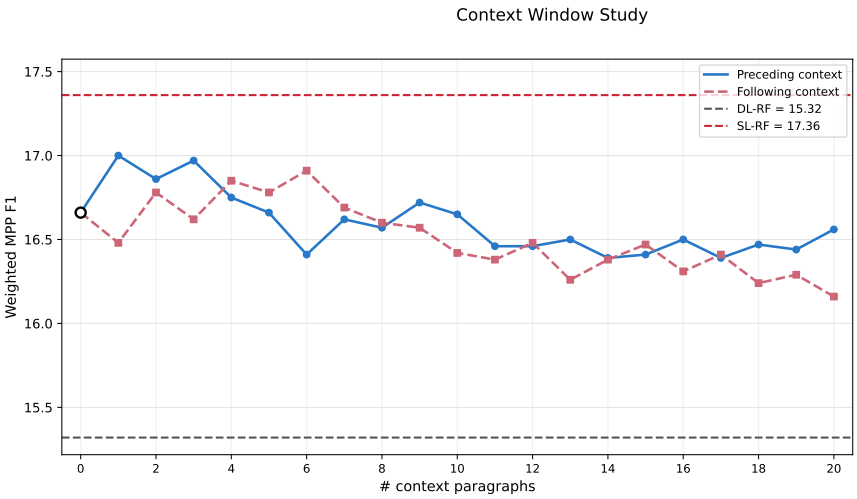
(b) EN→ES



N	DL-src (P)	DL-src (F)
0	87	87
2	215	233
4	299	338
6	350	416
8	378	471
10	391	510
12	394	542
14	394	572
16	394	597
18	394	616
20	394	633

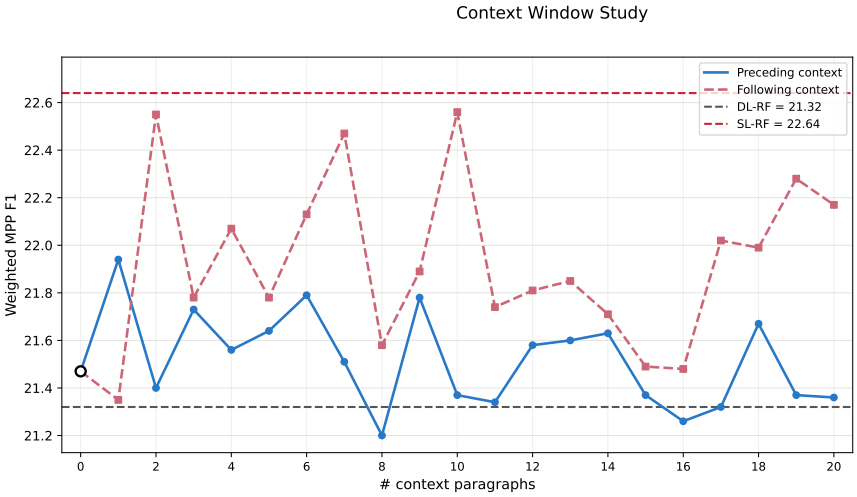
(c) JA→ZH

Figure 8: Context window study across WMT24 test set.



N	DL-src (P)	DL-src (F)
0	580	580
2	1470	1845
4	2198	2978
6	2888	3941
8	3544	4752
10	4165	5414
12	4747	5998
14	5291	6539
16	5798	7036
18	6267	7495
20	6697	7918

(a) EN→KO



N	DL-src (P)	DL-src (F)
0	186	186
2	437	516
4	625	775
6	785	966
8	923	1129
10	1050	1277
12	1172	1416
14	1285	1542
16	1385	1659
18	1474	1770
20	1553	1880

(b) JA→ZH

Figure 9: Context window study across WMT25 test set.