

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

Marina Iuliana Aur¹ Stefano Perrella¹ Roberto Navigli^{1,2}

¹Sapienza NLP Group, Sapienza University of Rome

²Babelscape, Rome, Italy

{aur,perrella,navigli}@diag.uniroma1.it

Abstract

Human annotators typically assess translation quality by considering the broader document context, which allows them to detect errors arising from discourse-level phenomena that may 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 present a controlled study of the impact of document context on LLM-based MT evaluation and its interaction with the availability of manually curated references. We consider several evaluation configurations, also exploring how performance varies when enabling the thinking inference mode and conducting task-specific fine-tuning. Our results show that the effect of document context differs substantially across test sets, and that input repetition itself can account for part of the apparent gains. Fine-tuning improves performance consistently and offers larger gains when document context is available, suggesting that task-specific adaptation may help models to handle document-level context more 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 relied on segment-level context, whereby text segments¹ are evaluated in isolation, independently of their surrounding context (Rei et al., 2020, 2022; Kocmi and Federmann, 2023a,b; Juraska et al., 2023, 2024; Guerreiro et al., 2024; Proietti et al., 2026). However, the lack of document-level context may prevent auto-evaluators from detecting errors arising from discourse-level phenomena, such

as anaphora, i.e., the use of pronouns or other expressions that refer to earlier words; lexical cohesion, i.e., consistency in the use of terminology, synonyms, hyponyms, etc.; and ellipsis, i.e., the omission of words that can be inferred from context (Wong and Kit, 2012; Bawden et al., 2018).

In human evaluation, Läubli et al. (2018) compared sentence-level and document-level evaluation, asking annotators to assess either isolated sentences or entire documents. They found that document-level evaluation revealed differences in translation quality that were less apparent at the sentence level. Following this work, subsequent editions of the Conference on Machine Translation (WMT) have increasingly incorporated document context into human evaluation, including protocols in which annotators assess individual segments while having access to the surrounding document (Barrault et al., 2019, 2020; Akhbardeh et al., 2021; Kocmi et al., 2022, 2023, 2024a, 2025).

In automatic evaluation, however, most approaches still operate using only segment-level context. The lack of document context can be partially mitigated by reference translations, which provide additional target-language evidence that may help resolve ambiguities in the source segment. However, reliance on reference translations may bias evaluation toward translations that resemble 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 relatively narrow range of applications, primarily the evaluation of MT systems on benchmarks for which references are available in advance. By contrast, reference-free auto-evaluators support a much broader 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., 2024a), and can

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

also serve as objectives for reinforcement learning-based fine-tuning of MT systems (Ramos et al., 2024; Xu et al., 2024; Zheng et al., 2025).

This work investigates the effects of incorporating document context into LLM-based MT evaluation and its interaction with reference translations. Specifically, we benchmark LLMs with and without thinking-enabled inference 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 explore the effects of fine-tuning across all evaluation settings. We release the code to reproduce our results at <https://github.com/SapienzaNLP/doc-context-mt-eval>.

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, 2023a,b; Juraska et al., 2023, 2024; Guerreiro et al., 2024).

Recent research has explored incorporating longer contexts into MT evaluation (Vernikos et al., 2022; Raunak et al., 2024; Wang et al., 2025; Domhan and Zhu, 2025). Vernikos et al. (2022) and Raunak et al. (2024) propose automatic procedures that extend existing segment-level MT evaluators to incorporate document context, while Wang et al. (2025) extend existing metrics to score longer documents through segmentation and alignment. However, the architectures underlying these evaluators have limited context windows – XLM-RoBERTa, for example, is restricted to 512 tokens (Conneau et al., 2020). Consequently, these approaches are limited in the amount of document context their underlying evaluators can process.

In contrast, Domhan and Zhu (2025) explore document-level MT evaluation using LLMs as judges, whose larger context windows make it possible to process substantially longer inputs. They show that LLM-based MT evaluation is not length-invariant, with automatic evaluators tending to detect fewer error spans as input texts become longer. Accordingly, their study focuses on mitigating evaluation degradation as input length increases, rather than systematically isolating the contribution of document context across evaluation settings with and without manually curated references.

2.2 Reference-free MT Evaluation

Representative reference-free evaluation approaches include 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-QE (Juraska et al., 2023, 2024, 2025). COMET-QE and MetricX-QE primarily perform score-level evaluation, while CometKiwi additionally supports word-level quality prediction. By contrast, MaTESe-QE (Perrella et al., 2022) and xCOMET-QE (Guerreiro et al., 2024) 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, a GPT-4-based (OpenAI et al., 2024b) auto-evaluator designed to detect translation errors following the Multidimensional Quality Metrics (MQM, Lommel et al., 2013; 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 Gemma 3 (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 auto-evaluators still typically underperform their reference-based counterparts (Freitag et al., 2022b, 2023, 2024; Lavie et al., 2025), suggesting that references provide useful evaluation signals beyond those that models can extract from the source sentence alone. It remains unclear, however, whether document-level context can compensate for the absence of references in span-level MT evaluation.

3 Methodology

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

Our work explores document context in 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 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 and how it interacts 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 our baseline setting, 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 allows us to analyze the impact of document-level context in the absence of human references, by providing the corresponding source and translation documents as context for segment-level evaluation. Although the model is still instructed to annotate only the translation segment, the translation document is included to expose target-side discourse phenomena, such as inconsistent translations across segments. DL-RB extends DL-RF by also including the human reference segment. This allows us to investigate whether human reference translations and document context provide complementary sources of information.

Lastly, document-level evaluation naturally repeats the source and translation segments, since they appear both as part of the document context and as the segments to be evaluated. Since preliminary results suggest that repetition may affect performance, we introduce an additional evaluation scenario, which we refer to as SL-RF_{rep}, to disentangle the effect of document-level context from that of repeated inputs. In SL-RF_{rep}, the evaluation is conducted exactly as in SL-RF, except that the source and translation segments are each presented twice in the input.

3.1 Error Detection

Approaches to automatic MT evaluation broadly fall into two categories, depending on their gran-

²Unlike them, we evaluate a single *segment* rather than a single *sentence* at a time, to align with the segment-level annotations of our test sets (Section 4.1).

	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” denote evaluation with segment-level and document-level context, respectively, while “RF” and “RB” stand for “reference-free” and “reference-based”, respectively.

ularity: score-level and span-level. Score-level evaluators assign a scalar quality score, which offers limited interpretability (Perrella et al., 2024a), as it conveys little information 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 each error a category and severity level. We therefore conduct our experiments at the span level, instructing LLMs to identify translation errors and classify them according to the MQM error taxonomy and one of three severity levels: Minor errors have a limited effect on meaning or readability, Major errors substantially affect meaning or readability, and Critical errors fundamentally alter the meaning or prevent comprehension³.

3.1.1 Meta-Evaluation

We measure performance using the MPP (Match with Partial overlap and Partial credit) measure⁴ 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. These per-span precision and recall values are then aggregated into their corpus-level counterparts used to compute the final *F*-score. We also extend MPP to incorporate error severity using the severity weighting strategy illustrated below. MQM error categories, instead, do not contribute to the MPP score, but we retain them to preserve the standard MQM annotation format.

Severity Weighting As it stands, MPP measures error detection performance irrespective of error severity. In previous work, error severity was incor-

³Although we ask our auto-evaluators to detect Critical errors, these are mapped to Major errors during meta-evaluation, because most test sets that we employ contain annotations featuring only Minor and Major errors.

⁴sted19/span-mt-metrics-eval

Test Set	LP	#docs	#segs	#spans	$\frac{\text{\#spans}}{\text{\#segs}}$
WMT24	EN→DE	60	6768	8173	1.21
	EN→ES	40	7224	2796	0.39
	JA→ZH	61	5936	4476	0.75
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 error spans per segment in the WMT test sets employed in our experiments.

porated into span-level meta-evaluation measures via the application of severity penalties (Blain et al., 2023; Zerva et al., 2024; Lavie et al., 2025), which discount the precision and recall obtained for 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 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.⁵ Formally, let us use the same notation as Perrella et al. (2026): predicted and ground-truth error spans are denoted by $\hat{s}$ and s , respectively, whereas $\hat{S}$ and S are sets of predicted and ground-truth error spans across an entire corpus. We define the set of all one-to-one matchings between $\hat{S}$ and S :

$$\mathcal{M} = \left\{ M \subseteq \hat{S} \times S \mid \begin{array}{l} \forall \hat{s} \in \hat{S}: |\{s: (\hat{s}, s) \in M\}| \leq 1 \\ \forall s \in S: |\{\hat{s}: (\hat{s}, s) \in M\}| \leq 1 \end{array} \right\}.$$

We then select $M^* \in \mathcal{M}$ as the matching that maximizes the F -score. Given a matched pair $(\hat{s}, s) \in M^*$, per-span precision and recall – based on the proportion of overlapping characters – are $P_{\text{MPP}}(\hat{s}, s)$ and $R_{\text{MPP}}(\hat{s}, s)$, respectively. MPP derives corpus-level precision and recall as follows:

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

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

Unlike 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, an unmatched Minor prediction should be penalized less than an unmatched Major prediction. Finally, as with the severity penalties employed in previous work, severity mismatches between overlapping errors are penalized, because per-span precision and recall are rescaled using the minimum of the involved spans’ severity weights.

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 in WMT campaigns have been shown to identify more errors than ESA annotators (Kocmi et al., 2024b), making MQM data more suitable for error detection. Furthermore, we found that more than half (63.0% in WMT24 and 51.4% in WMT25) of error-free ESA translations were assigned final quality scores lower than 100 (perfect score), suggesting that ESA data might be noisier when used for span-level MT evaluation, as opposed to score-based evaluation.

The WMT24 and WMT25 MQM test sets cover four domains: literary, news, social, and speech. We retain literary, news, and social data, while excluding speech, for which document and segment boundaries coincide and no additional document context is therefore available beyond the evaluated

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

Test set	Model	Think	Evaluation Scenarios				Repetition Control
			SL-RF	SL-RB	DL-RF	DL-RB	SL-RF _{rep}
WMT24 EN→DE EN→ES JA→ZH	<i>Baseline</i>						
	xCOMET-XL	×	9.58	<u>11.78</u>	–	–	–
	<i>Qwen3</i>						
	Qwen3-8B	×	15.46	16.09	15.98	16.34	<u>17.26</u>
	Qwen3-8B	✓	20.19	18.95	<u>21.73</u>	20.25	<u>20.95</u>
	Qwen3-32B	×	20.97	20.11	22.41	20.83	<u>22.69</u>
	Qwen3-32B	✓	23.21	20.87	24.54	21.86	23.62
	<i>Gemma-4</i>						
	Gemma-4-E4B	×	21.65	21.44	22.17	21.74	22.01
	Gemma-4-E4B	✓	20.85	20.32	<u>22.52</u>	<u>22.58</u>	21.96
WMT25 EN→KO JA→ZH	<i>Qwen3 Fine-tuned</i>						
	Qwen3-8B-fine-tuned	×	16.75	19.68	20.16	<u>21.69</u>	–
	<i>Baseline</i>						
	xCOMET-XL	×	12.93	<u>15.40</u>	–	–	–
	<i>Qwen3</i>						
	Qwen3-8B	×	14.79	<u>18.96</u>	12.33	15.20	14.90
	Qwen3-8B	✓	14.60	<u>19.63</u>	10.95	14.70	13.74
	Qwen3-32B	×	20.61	23.84	19.05	21.26	20.44
	Qwen3-32B	✓	13.81	<u>20.96</u>	12.49	19.17	13.97
	<i>Gemma-4</i>						
	Gemma-4-E4B	×	17.02	<u>21.01</u>	11.28	12.97	13.95
	Gemma-4-E4B	✓	13.15	<u>18.09</u>	11.21	13.77	12.82
	<i>Qwen3 Fine-tuned</i>						
	Qwen3-8B-fine-tuned	×	17.36	<u>19.12</u>	17.61	18.74	–

Table 3: Weighted micro-averaged MPP F -score on the WMT24 and WMT25 MQM test sets. The first four columns report the main evaluation scenarios. The repetition control, SL-RF_{rep}, repeats the evaluated source and translation segments without adding document context. The “Think” column indicates whether thinking is enabled. The best result per test set among the main evaluation scenarios is in bold, while the best result per model is underlined.

segment. We report statistics about the WMT test sets and language pairs (LPs) employed in this work in Table 2 and Appendix A.

4.2 Prompt Templates

The prompt templates used to implement the different evaluation scenarios, including the SL-RF_{rep} repetition control, are reported in Appendix B. All scenarios share the same system prompt, error taxonomy, severity definitions, and output format, differing only in their inputs and scenario-specific instructions.

Following GemSpanEval (Juraska et al., 2025), we also ask the model to return an extended span containing surrounding context, which is used to disambiguate repeated occurrences of the same minimal span (e.g., when the same word or phrase occurs multiple times in a segment). When a minimal span occurs multiple times in the translation, we use the corresponding extended span to disambiguate the intended occurrence before matching predicted and ground-truth errors with MPP.

4.3 Models

Our auto-evaluators consist of three instruction-tuned LLMs spanning two model families and different scales: Qwen3-8B, Qwen3-32B (Yang et al., 2025), and Gemma-4-E4B (Team et al., 2026), with context windows of 32K, 32K, and 128K, respectively. 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 allows us to compare MT error detection performance under standard and thinking-enabled inference. Additional details, including decoding parameters and thinking configuration, are provided in Appendix C.

4.4 Fine-Tuning

To the best of our knowledge, previous fine-tuning approaches for span-level MT evaluation have largely focused on segment-level inputs. Recently, Domhan and Zhu (2025) explored fine-tuning GPT-4o (OpenAI et al., 2024a) to conduct evaluation at the segment, document, and multi-document lev-

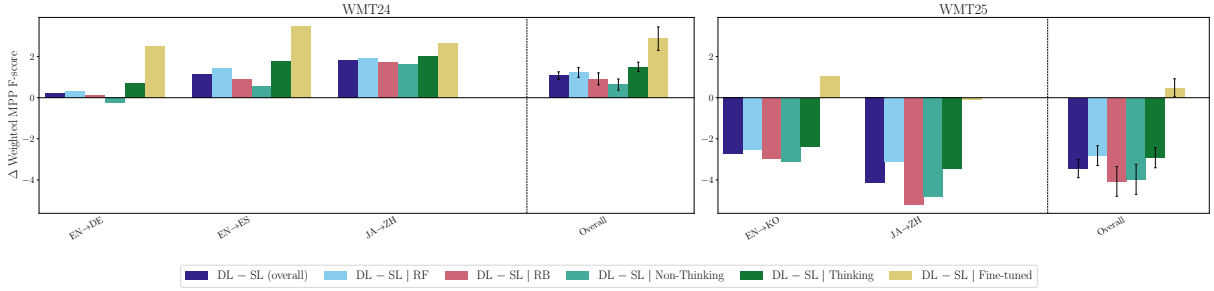


Figure 1: Average effect of adding document context on weighted MPP F -score across language pairs and test sets. “DL - SL (overall)” reports the difference between all configurations extended with document context and segment-level configurations. Other entries of the legend correspond to the same effect, but computed across configurations that share a fixed evaluation dimension, e.g., “DL - SL | RF” only includes reference-free evaluations. All effects are computed across non-fine-tuned auto-evaluators, except for “DL - SL | Fine-tuned”, which isolates the effect of document context across fine-tuned auto-evaluators. Error bars on the overall estimates indicate 95% confidence intervals obtained with paired document-level bootstrap resampling. Full estimates and bootstrap details are reported in Appendix F.

els. Instead, we investigate fine-tuning models to exploit broader document context while evaluating a fixed segment at a time.

Specifically, we fine-tune Qwen3-8B four times, once for each of the four evaluation scenarios in Table 1. For each scenario, training data are drawn from the WMT22 and WMT23 MQM test sets and comprise 89,480 examples across four language pairs. Further details about our fine-tuning data and experimental setup are available in Appendix D. Notably, EN→DE is the only translation direction shared between the training and evaluation sets.

5 Results

Table 3 reports the weighted MPP F -score on the WMT24 and WMT25 MQM test sets across models and evaluation scenarios,⁶ while a breakdown of per-language-pair results is provided in Tables 12 and 13 in Appendix E. We also report the performance of xCOMET-XL (Guerreiro et al., 2024) to contextualize absolute performance relative to an existing span-level MT auto-evaluator. The best overall configuration on WMT24 is **DL-RF** with Qwen3-32B and thinking enabled (24.54), while on WMT25 the best result is achieved by **SL-RB** with Qwen3-32B and thinking disabled (23.84).

Figure 1 summarizes the effect of document context on weighted MPP F -score for each language pair and overall, across WMT24 and WMT25, and also reports the effect of document context while

holding other evaluation dimensions fixed, namely reference availability, thinking-enabled inference, and task-specific fine-tuning. As shown in the figure, incorporating document context into the evaluation has mixed effects depending on the test set. Specifically, on WMT24, most bars are positive, suggesting that adding document context may improve evaluation performance, whereas the opposite pattern emerges on WMT25. Incorporating document-level context improves weighted MPP F -score for fine-tuned models across both test sets, with the only exception of WMT25 JA→ZH, where the effect is slightly negative. Moreover, we note that, even on WMT24, the gains brought about by document context for fine-tuned models are generally larger than those observed in the other evaluation scenarios. Lastly, we note that document context is more beneficial – or, on WMT25, less detrimental – when manually curated references are unavailable, suggesting that it may partially compensate for their absence.

In the following sections, we examine these results in greater detail. We first provide an explanation for why the effects are reversed across test sets (Section 5.1), then investigate the impact of repeating the input segments twice in the prompt (Section 5.2), and finally delve deeper into the performance of our fine-tuned models (Section 5.3). Section 6 then investigates how performance changes with the amount of document context.

5.1 Understanding Test-Set Differences

We identify two main differences between the WMT24 and WMT25 test sets: (i) source docu-

⁶As recommended by Perrella et al. (2026), we only report MPP results micro-averaged across test set samples, since macro-averaging was found to be biased toward auto-evaluators that detect fewer error spans.

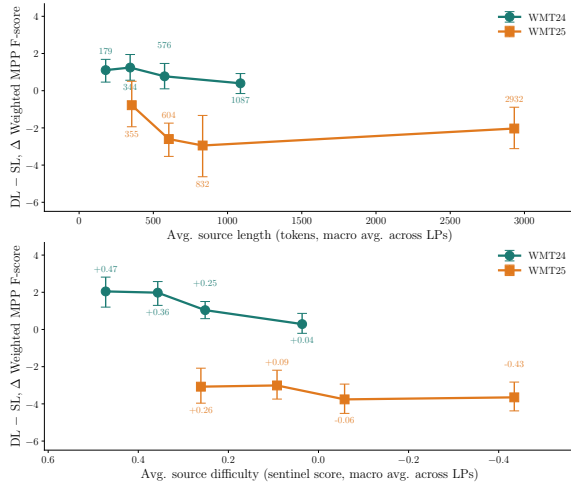


Figure 2: Document-context effect (DL - SL) by source document length (top) and segment difficulty (bottom), split into four per-language-pair quartiles.

ment length and (ii) source segment difficulty. Furthermore, we show that not only document context but also reference-based evaluation and thinking-enabled inference have contrasting effects across the two test sets.

The Impact of Source Document Length Documents and segments in WMT25 are substantially longer than those in WMT24 (see Table 4 in Appendix A). The difference is particularly large for EN→KO, whose source documents are about 2.6 times longer than those in WMT24 EN→ES and EN→DE; WMT25 JA→ZH documents are about 1.7 times longer than their WMT24 counterparts.

To study the role of document length, we tokenize the source documents using the Qwen3-32B tokenizer and partition them into four per-language-pair length quartiles based on token count. We then compute the effect of document context within each quartile. As shown in Figure 2, the two test sets cover substantially different length ranges, with even the shortest WMT25 documents being relatively long compared with those in WMT24. In both test sets, gains generally decrease with document length, suggesting that longer documents make context harder to exploit. However, on WMT25, document context is harmful even for quartiles whose length ranges overlap with lengths observed in some WMT24 documents, where document context improves performance. Thus, while longer inputs may make context harder to exploit, document length alone does not fully explain the difference between the two test sets.

The Impact of Source Segment Difficulty Another important distinction between the two test sets is their sampling strategy. As reported by Kocmi et al. (2025), WMT25 documents were subsampled from a larger pool using the sentinel-src-25⁷ difficulty estimator (Proietti et al., 2025), which builds on sentinel-src (Perrella et al., 2024b), to select more difficult documents for inclusion in the test set. In contrast, documents in the WMT24 test set were randomly sampled. We therefore examine whether the effect of document context varies with source difficulty. We apply sentinel-src-25 to all source segments to quantify their difficulty, with higher scores indicating easier segments. We then partition the source segments into four per-language-pair difficulty quartiles and compute the document-context effect within each quartile.

As shown in Figure 2, the document-context effect decreases as source difficulty increases in both test sets. As with document length, source difficulty alone does not fully explain the differences between the two test sets. Nonetheless, taken together, document length and source difficulty partly explain the contrasting effect of document context across WMT24 and WMT25.

The Effect of Human References From the results in Table 3 we calculate that human references show a small negative overall effect on WMT24 (-0.42, 95% CI [-0.92, 0.10]) but a positive overall effect on WMT25 (+3.65, 95% CI [2.84, 4.64]). On WMT24, their effect is small or negative, particularly for EN→ES, where the reference quality is lower (Freitag et al., 2024, Table 5). In contrast, references substantially improve performance on WMT25, suggesting that they may be especially useful on more difficult test sets, in line with Lavie et al. (2025).

The Effect of Thinking In contrast to reference availability, thinking-enabled inference improves performance on WMT24 (+1.89, 95% CI [1.43, 2.34]) but reduces it on WMT25 (-2.15, 95% CI [-3.07, -1.39]). This difference may arise because thinking consistently reduces the number of predicted error spans, which may be particularly detrimental on WMT25, whose segments contain more errors on average. Document context likewise reduces the number of predicted spans, whereas manually curated references increase it (Figure 5

⁷Prosho/sentinel-src-25

in Appendix G). These effects on span counts may partly explain the differing results across WMT24 and WMT25.

5.2 Disentangling Document Context from Input Repetition

As mentioned in Section 3, one important aspect of settings with access to document context is that the evaluated source and translation segments are each presented to the model twice: once explicitly as the segments to be evaluated, and once as part of their corresponding full documents. This raises the question of whether the gains brought about by document context on WMT24 stem from the models effectively leveraging the additional context or, at least in part, from input repetition. Indeed, [Leviathan et al. \(2025\)](#) show that repeating the input prompt can improve performance across seven benchmarks.

To investigate this, we include a repetition control scenario in Table 3, denoted `SL-RFrep`, which uses the same prompt as `SL-RF` but presents both the source and translation segments twice. Then, we measure the effect of `DL-RF` - `SL-RFrep`, i.e., the benefit of adding document context in reference-free scenarios over simply repeating the source and translation segments in the input prompt, yielding $+0.14$ (95% CI $[-0.34, 0.64]$) on WMT24, and -2.09 (95% CI $[-2.56, -1.36]$) on WMT25. These results suggest that most of the gains brought about by document context on WMT24, especially in the reference-free setting, may actually be due to input repetition. On WMT25, instead, the effect of adding document context is already negative, and these results further show that document context decreases performance even compared with input repetition. On WMT24, our results also align with the pattern reported by [Leviathan et al. \(2025\)](#): prompt repetition yields larger gains when thinking is disabled, while its effect is mostly neutral to slightly positive when thinking is enabled. This pattern, however, does not extend to WMT25.

5.3 Fine-Tuning MT Evaluation with Document Context

As shown in Table 3, task-specific fine-tuning improves performance across most evaluation scenarios and test sets. Figure 3 further shows that these gains are generally larger in document-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

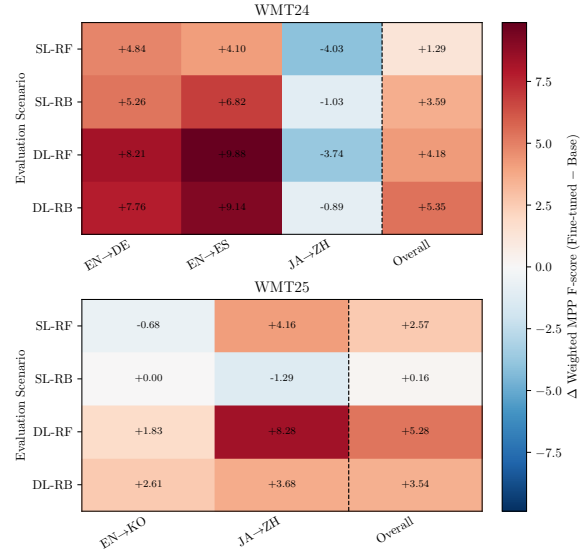


Figure 3: Impact of fine-tuning on weighted MPP F -score across evaluation scenarios, test sets, and language pairs. Values report the difference between fine-tuned Qwen3-8B and the corresponding base Qwen3-8B configuration with thinking disabled. Positive values indicate gains from fine-tuning, while negative values indicate drops.

notably negative impact on WMT25 performance for the base models. Nonetheless, fine-tuning improves performance in both document-level settings (`DL-RF` and `DL-RB`), with larger gains than in their segment-level counterparts (`SL-RF` and `SL-RB`). Furthermore, despite `SL-RB` remaining the strongest fine-tuned configuration on WMT25, fine-tuning substantially narrows the gap between segment- and document-level evaluation: compared with their segment-level counterparts, the gains from fine-tuning are $+2.71$ points larger for `DL-RF` (95% CI $[1.28, 4.05]$) and $+3.37$ points larger for `DL-RB` (95% CI $[2.01, 4.93]$). This suggests that task-specific adaptation can help models better handle the additional complexity introduced by document-level inputs.

6 When Is Document Context Useful?

The results in Section 5.2 show that the apparent advantage of document-level evaluation on WMT24 is largely explained by input repetition rather than by the surrounding document context itself. Moreover, the analysis in Section 5.1 suggests that longer and more difficult inputs may make document context more difficult to exploit, leading to performance degradation on WMT25. Nonetheless, the fine-tuning results in Section 5.3 indicate that

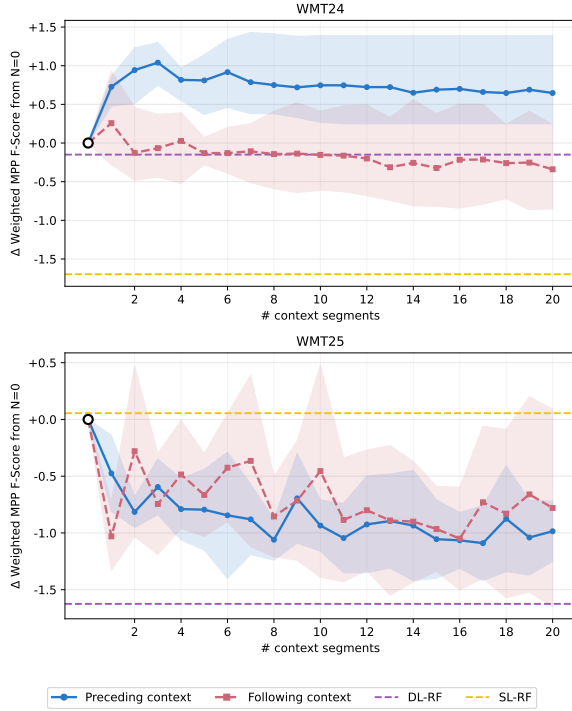


Figure 4: Average effect of adding preceding and following context segments on weighted MPP F -score, relative to `SL-RFrep`. Curves show the mean across language pairs within each test set; shaded bands indicate the per-language-pair range.

document context can still provide a useful evaluation signal. We therefore ask whether the broader document context can be better exploited when the amount of additional context is controlled.

To investigate this, we vary the number and position of additional context segments provided to Qwen3-32B, the strongest model in our experiments, and measure weighted MPP F -score as preceding or following source and translation segments are added as context. All evaluations use the `DL-RF` prompt template with thinking disabled. We denote by N the number of additional context segments. At $N = 0$, the `DL-RF` template reduces to the `SL-RFrep` scenario, where the source and translation segments are presented twice. Thus, $N = 0$ serves as the baseline for measuring the contribution of additional context. Figure 4 reports results aggregated across translation directions, while a per-language-pair breakdown is provided in Appendix H. We also report `DL-RF` and `SL-RF` for comparison.

On WMT24, evaluation performance increases when preceding segments are provided as additional context. However, the gains saturate quickly, at roughly three preceding segments, beyond which

additional context leaves performance unchanged or slightly degrades it. By contrast, following context generally provides little or no benefit, with performance tending to decrease as more context is added. This result extends the findings of Section 5.2: document context can improve evaluation beyond the repetition effect, but the benefit is concentrated in a small amount of preceding context and diminishes as more context is added.

On WMT25, additional context provides no consistent benefit and generally reduces performance. We hypothesize that this stems from the considerations discussed above: greater source difficulty and longer inputs may make it more challenging for the model to effectively leverage the additional context on WMT25, or even to handle repeated inputs. Indeed, `SL-RFrep` slightly underperforms `SL-RF` on this test set, and `DL-RF` performs substantially worse than both.

7 Conclusion

This work investigates how document context affects LLM-based span-level MT evaluation, finding that its impact varies substantially across test sets. On WMT24, adding document context improves evaluation performance, although most of the benefit can be explained by the confounding effect of input repetition. Nonetheless, we find that providing only a small amount of preceding context can yield benefits beyond those attributable to input repetition and further improve performance. On WMT25, by contrast, both repetition and additional document context reduce performance. We attribute these reversed effects to the longer and more difficult-to-translate source texts in WMT25.

Regarding the relationship between document context and the availability of reference translations, document context appears to be more beneficial when references are unavailable, suggesting that it may partially compensate for their absence.

Lastly, we find that task-specific fine-tuning consistently improves performance across both test sets, with larger gains in document-level settings. This suggests that task-specific adaptation may help models handle document-level inputs more effectively. We recommend that future work investigate whether these fine-tuning gains extend to larger models and broader evaluation settings, potentially enabling state-of-the-art MT evaluation through more effective 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 all 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 constrained by the availability of MQM annotations in the WMT24 and WMT25 test sets. We restrict our analysis to these two test sets, excluding older MQM data to reduce the risk of evaluating on material that may have been included in the LLMs’ pre-training corpora.

Fine-Tuning Scope Our fine-tuning experiments are limited to Qwen3-8B, as their purpose is diagnostic. Our results suggest that document-level information carries a meaningful signal and that Qwen3-8B can learn to exploit it through task-specific fine-tuning. In particular, fine-tuning yields larger gains in document-level settings, suggesting that task-specific adaptation helps the model make more effective use of the additional context. Whether these gains generalize to larger models or other model families remains an open question for future work.

Acknowledgments

The authors acknowledge ISCRA for granting this project access to the LEONARDO supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CINECA (Italy). Marina Iuliana Aur and Roberto Navigli gratefully acknowledge the support of the IT4LIA AI Factory project, grant agreement no. 101234224.

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 M. 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-jussa, 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-jussa, 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.

Frederic Blain, Chrysoula Zerva, Ricardo Rei, Nuno M. Guerreiro, Diptesh Kanojia, José G. C. de Souza, Beatriz Silva, Tânia Vaz, Yan Jingxuan, Fatemeh Azadi, Constantin Orasan, and André Martins. 2023. [Findings of the WMT 2023 shared task on quality estimation](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 629–653, Singapore. 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 7929–7947, 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 Lakounga, 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.
- Philipp Koehn. 2004. [Statistical significance tests for machine translation evaluation](#). In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. [Efficient memory management for large language model serving with pagedattention](#). *Preprint*, arXiv:2309.06180.
- Samuel Lübl, 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 Sindhuja, 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 Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2013. [Multidimensional quality metrics: a flexible system for assessing translation quality](#). In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.
- 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, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024a. [Gpt-4o system card](#). *Preprint*, arXiv:2410.21276.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024b. [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 Hugué Cabot, Edoardo Barba, and Roberto Navigli. 2024a. [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, Edoardo Barba, and Roberto Navigli. 2024b. [Guardians of the machine translation meta-evaluation: Sentinel metrics fall in!](#) In *Proceedings of the 62nd Annual Meeting of the Association*

- for *Computational Linguistics (Volume 1: Long Papers)*, pages 16216–16244, Bangkok, Thailand. 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, Roman Grundkiewicz, and Matt Post. 2026. [PEAR: Pairwise evaluation for automatic relative scoring in machine translation](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 42189–42207, San Diego, California, United States. 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, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, and 304 others. 2026. [Gemma 4 technical report](#). Preprint, arXiv:2607.02770.
- 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.
- 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 32323–32339, 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.

Chrysoula Zerva, Frederic Blain, José G. C. De Souza, Diptesh Kanojia, Sourabh Deoghare, Nuno M. Guerreiro, Giuseppe Attanasio, Ricardo Rei, Constantin Orasan, Matteo Negri, Marco Turchi, Rajen Chatterjee, Pushpak Bhattacharyya, Markus Freitag, and André Martins. 2024. [Findings of the quality estimation shared task at WMT 2024: Are LLMs closing the gap in QE?](#) In *Proceedings of the Ninth Conference on Machine Translation*, pages 82–109, Miami, Florida, USA. 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 Test Set Statistics

Table 4 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 their language pairs, but also in their structure.

B Evaluation Prompts

Table 5 reports the unified prompt template used across all main evaluation scenarios presented in Section 3. The scenarios share the same system prompt, error taxonomy, severity definitions, and output format, while differing only in the input fields and task instructions required by each setting. Additionally, Table 6 reports the prompt template used for the `SL-RFrep` repetition control, presented in Section 3 and discussed in Section 5.2. To implement input repetition, the source segment and its translation are presented twice before the task instructions.

C Inference Configuration

We run the models described in Section 4.3 using vLLM (Kwon et al., 2023). For non-thinking inference, we use greedy decoding. For thinking-enabled inference, we first tried using greedy decoding as well, but noticed significant inconsistencies in the models’ outputs. We therefore use the sampling configurations recommended for each model by their respective authors. Specifically, Qwen3 uses temperature 0.6, top-p 0.95, and top-k 20, whereas Gemma-4-E4B uses temperature 1.0, top-p 0.95, and top-k 64. We use a random seed of 42 for sampling.

D Fine-Tuning Setup

This appendix complements the fine-tuning experiments presented in Section 5.3. Specifically, Tables 7 and 8 report general statistics and character-length statistics for the training data, while Table 9 summarizes the training configuration.

We fine-tune the model separately for each evaluation scenario, using the same input structure and output format as at inference time. We remove individual examples where the segment coincides with the entire document, normalize error categories, and exclude all examples from domains not considered in our evaluation, i.e., domains in which the evaluated segment coincides with the entire

Test set	LP	Stat.	Segment Length	Document Length
WMT24	EN→DE	avg	154.2	2368.7
		std	158.5	1700.7
		med	91.5	1843.5
		max	886.0	5786.0
	EN→ES	avg	186.4	2428.3
		std	177.9	1636.1
		med	114.5	1868.5
		max	967.0	5667.0
	JA→ZH	avg	87.4	776.0
		std	78.1	489.2
		med	67.5	664.0
		max	531.0	1701.0
WMT25	EN→KO	avg	575.0	6233.5
		std	220.9	7138.9
		med	620.0	4633.5
		max	1057.0	28560.0
	JA→ZH	avg	181.3	1313.8
		std	139.3	2159.0
		med	149.0	573.0
		max	600.0	11272.0

Table 4: 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, computed over unique segments and unique documents, respectively.

document. We then remove problematic annotations in which the same error span occurs at least five times, to prevent degenerate repetition during generation. After preprocessing, the training set contains 89,480 examples for each evaluation scenario.

E Detailed Evaluation Results

This appendix provides additional results that complement the aggregate weighted MPP F -score results reported in Table 3 in Section 5. Specifically, Tables 10 and 11 report weighted MPP precision and recall scores, respectively. In addition, Tables 12 and 13 present weighted MPP F -score broken down by language pair and evaluation scenario, showing whether the main trends hold across translation directions.

F Bootstrap Analysis

We assess uncertainty in the marginal-effect estimates presented in Section 5 using paired bootstrap resampling (Koehn, 2004). For each test set, we construct 10,000 bootstrap samples by resampling documents with replacement within each language pair, while keeping the original number of documents per language pair. Within each replicate, the same resampled documents are used for all model and evaluation configurations.

For each configuration, we recompute the corpus-level severity-weighted MPP precision and recall by summing the per-span numerators and denominators over the resampled documents before taking their ratio. For each factor, we compare configurations that differ only in that factor and average the resulting differences. The fine-tuning effect is computed for Qwen3-8B across the four evaluation scenarios. The 95% confidence intervals correspond to the 2.5th and 97.5th percentiles of the resulting bootstrap distributions.

Table 14 reports the document-context estimates and confidence intervals shown in Figure 1.

G Predicted Error Span Analysis

Figure 5 reports how document context, reference availability, and thinking affect the number of predicted error spans per segment.

H Amount and Position of Additional Context

This appendix provides the per-language-pair results for the analysis described in Section 6. Figures 6 and 7 show how performance varies as a function of the size and position of the document context provided to the model on WMT24 and WMT25, respectively.

System prompt

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

User prompt

{source_lang} Full Source Document:

```{source\_document}```

{target\_lang} Full Translation Document:

```{translation\_document}```

{source_lang} Source Segment:

```{source\_segment}```

{target\_lang} Reference Segment:

```{reference\_segment}```

{target_lang} Translation Segment:

```{translation\_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 and the reference 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 5: Unified prompt template used across all main evaluation scenarios. Gray-shaded fields are included only when required by the selected scenario: reference segments in RB settings, and source and 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.

### System prompt

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

### User prompt

{source\_lang} Source Segment:  
```{source\_segment}```

{target_lang} Translation Segment:
```{translation\_segment}```

{source\_lang} Source Segment:  
```{source\_segment}```

{target_lang} Translation Segment:
```{translation\_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.

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 6: Prompt template for `SL-RFrep`. It follows the same structure as `SL-RF`, but presents the source and translation segments twice before the evaluation instructions.

Year	LP	#docs	#segs	#spans	$\frac{\text{\#spans}}{\text{\#segs}}$
WMT22	EN→DE	179	15406	14361	0.93
	EN→RU	176	15065	21450	1.42
	EN→ZH	179	22487	30858	1.37
	ZH→EN	185	25925	29987	1.16
WMT23	EN→DE	83	3385	8186	2.42
	ZH→EN	52	7212	16161	2.24

Table 7: Number of evaluated documents, segments, error spans, and average error spans per segment in the fine-tuning training data (WMT22/WMT23), after pre-processing.

Year	LP	Stat.	Segment Length	Document Length
WMT22	EN→DE	avg	94.8	1099.0
		std	61.2	1001.9
		med	80.0	690.0
		max	519.0	4418.0
	EN→RU	avg	94.8	1113.0
		std	61.2	1004.6
		med	80.0	703.5
		max	519.0	4418.0
	EN→ZH	avg	97.3	1099.0
		std	60.3	1001.9
		med	84.0	690.0
		max	519.0	4418.0
	ZH→EN	avg	40.8	414.1
		std	31.7	442.2
		med	34.0	141.0
		max	299.0	1729.0
WMT23	EN→DE	avg	258.2	1154.8
		std	258.0	1001.0
		med	151.0	512.0
		max	1522.0	4316.0
	ZH→EN	avg	48.5	1099.5
		std	34.6	628.3
		med	41.0	1032.5
		max	275.0	3105.0

Table 8: Character-length statistics for the fine-tuning training data (WMT22/WMT23) by language pair. Segment Length and Document Length denote source-side segment- and document-level lengths in characters, computed over unique segments and unique documents, respectively.

Training Configuration	Value
Learning rate	$1 \times 10^{-4}$
Weight decay	0.01
Effective batch size	32
Mini batch size	8
Gradient accumulation steps	1
Number of nodes	1
GPUs per node	4
Epochs	1
Optimizer	AdamW
LR scheduler	Linear
Max sequence length	8,192
LoRA Configuration	Value
Rank ( $r$ )	16
Alpha ( $\alpha$ )	32
Dropout	0.05

Table 9: Fine-tuning configuration.

Test set	Model	Think	Evaluation Scenarios			
			SL-RF	SL-RB	DL-RF	DL-RB
WMT24 EN→DE EN→ES JA→ZH	<i>Baseline</i>					
	xCOMET-XL	×	5.54	<u>6.93</u>	-	-
	<i>Qwen3</i>					
	Qwen3-8B	×	10.14	9.75	<u>10.56</u>	10.32
	Qwen3-8B	✓	17.03	12.92	<u>20.44</u>	15.26
	Qwen3-32B	×	14.42	12.70	<u>16.58</u>	13.87
	Qwen3-32B	✓	20.97	14.18	<b><u>25.31</u></b>	15.53
	<i>Gemma-4</i>					
	Gemma-4-E4B	×	17.24	14.67	<u>22.45</u>	18.59
	Gemma-4-E4B	✓	19.60	15.17	<u>24.66</u>	19.59
	<i>Qwen3 Fine-tuned</i>					
	Qwen3-8B-fine-tuned	×	10.73	13.11	14.02	<u>15.30</u>
WMT25 EN→KO JA→ZH	<i>Baseline</i>					
	xCOMET-XL	×	8.78	<u>11.17</u>	-	-
	<i>Qwen3</i>					
	Qwen3-8B	×	12.84	<u>13.96</u>	10.98	12.41
	Qwen3-8B	✓	21.31	<u>20.04</u>	<u>24.46</u>	21.83
	Qwen3-32B	×	18.23	<u>19.89</u>	17.94	19.12
	Qwen3-32B	✓	25.35	23.01	<b><u>30.23</u></b>	23.87
	<i>Gemma-4</i>					
	Gemma-4-E4B	×	21.70	19.57	<u>26.38</u>	22.68
	Gemma-4-E4B	✓	24.00	22.12	<u>29.68</u>	26.72
	<i>Qwen3 Fine-tuned</i>					
	Qwen3-8B-fine-tuned	×	11.79	14.10	14.98	<u>17.03</u>

Table 10: 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 thinking enabled. The best result per test set is in bold, while the best result per model is underlined.

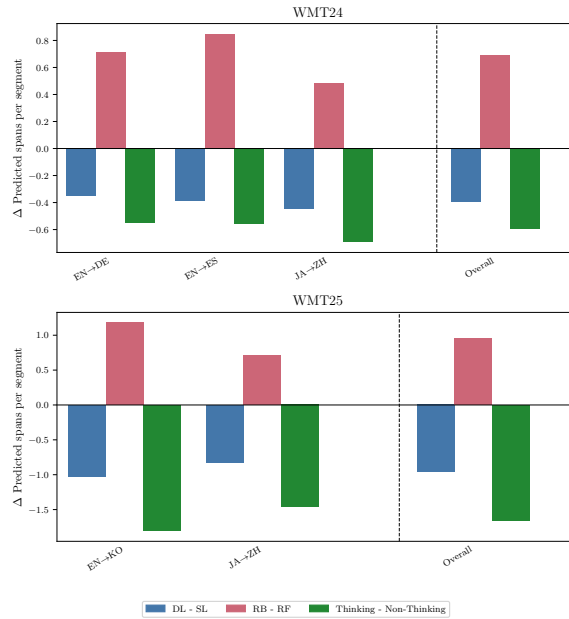


Figure 5: Average marginal effects on the number of predicted error spans per segment across language pairs and test sets. Positive values indicate more predicted spans, while negative values indicate fewer.

Test set	Model	Think	Evaluation Scenarios			
			SL-RF	SL-RB	DL-RF	DL-RB
WMT24 EN→DE EN→ES JA→ZH	<i>Baseline</i>					
	xCOMET-XL	×	35.47	<u>39.04</u>	-	-
	<i>Qwen3</i>					
	Qwen3-8B	×	32.51	<u>45.98</u>	32.85	39.19
	Qwen3-8B	✓	24.78	<u>35.53</u>	23.20	30.08
	Qwen3-32B	×	38.38	<b>48.32</b>	34.53	41.80
	Qwen3-32B	✓	25.97	<u>39.45</u>	23.81	36.91
	<i>Gemma-4</i>					
	Gemma-4-E4B	×	29.09	<u>39.77</u>	21.91	26.17
	Gemma-4-E4B	✓	22.27	<u>30.78</u>	20.72	26.64
	<i>Qwen3 Fine-tuned</i>					
	Qwen3-8B-fine-tuned	×	38.13	<u>39.45</u>	35.82	37.24
WMT25 EN→KO JA→ZH	<i>Baseline</i>					
	xCOMET-XL	×	24.52	<u>24.78</u>	-	-
	<i>Qwen3</i>					
	Qwen3-8B	×	17.44	<u>29.51</u>	14.04	19.62
	Qwen3-8B	✓	11.11	<u>19.23</u>	7.05	11.09
	Qwen3-32B	×	23.70	<u>29.74</u>	20.31	23.93
	Qwen3-32B	✓	9.49	<u>19.24</u>	7.87	16.01
	<i>Gemma-4</i>					
	Gemma-4-E4B	×	14.00	<u>22.68</u>	7.17	9.09
	Gemma-4-E4B	✓	9.06	<u>15.31</u>	6.91	9.27
	<i>Qwen3 Fine-tuned</i>					
	Qwen3-8B-fine-tuned	×	<b>32.94</b>	29.70	21.35	20.83

Table 11: 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 thinking 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	<i>Baseline</i>					
	xCOMET-XL	×	15.63	<u>17.59</u>	-	-
	<i>Qwen3</i>					
	Qwen3-8B	×	17.48	<u>19.20</u>	17.65	18.19
	Qwen3-8B	✓	20.34	<u>21.22</u>	21.08	21.13
	Qwen3-32B	×	24.18	23.82	<u>25.19</u>	24.30
	Qwen3-32B	✓	23.76	23.33	23.63	<u>25.05</u>
	<i>Gemma-4</i>					
	Gemma-4-E4B	×	21.83	<u>23.59</u>	21.21	22.07
	Gemma-4-E4B	✓	20.43	22.06	21.22	<u>23.25</u>
	<i>Qwen3 Fine-tuned</i>					
	Qwen3-8B-fine-tuned	×	22.32	24.46	25.86	<b><u>25.95</u></b>
EN→ES	<i>Baseline</i>					
	xCOMET-XL	×	5.62	<u>7.58</u>	-	-
	<i>Qwen3</i>					
	Qwen3-8B	×	<u>9.32</u>	9.08	8.81	8.42
	Qwen3-8B	✓	15.69	11.24	<u>17.00</u>	12.46
	Qwen3-32B	×	15.93	13.13	<u>17.95</u>	13.62
	Qwen3-32B	✓	20.79	13.95	<b><u>23.01</u></b>	14.23
	<i>Gemma-4</i>					
	Gemma-4-E4B	×	16.89	13.32	<u>17.74</u>	14.39
	Gemma-4-E4B	✓	18.23	13.66	<u>20.88</u>	16.59
	<i>Qwen3 Fine-tuned</i>					
	Qwen3-8B-fine-tuned	×	13.42	15.90	<u>18.69</u>	17.56
JA→ZH	<i>Baseline</i>					
	xCOMET-XL	×	9.88	<u>12.08</u>	-	-
	<i>Qwen3</i>					
	Qwen3-8B	×	18.25	18.53	20.46	<u>21.17</u>
	Qwen3-8B	✓	22.50	22.46	24.66	<u>25.02</u>
	Qwen3-32B	×	21.02	21.51	<u>22.63</u>	22.29
	Qwen3-32B	✓	23.69	23.56	<u>25.91</u>	24.40
	<i>Gemma-4</i>					
	Gemma-4-E4B	×	24.06	25.15	25.44	<b><u>26.25</u></b>
	Gemma-4-E4B	✓	22.52	23.02	24.51	<u>25.46</u>
	<i>Qwen3 Fine-tuned</i>					
	Qwen3-8B-fine-tuned	×	14.22	17.50	16.72	<u>20.28</u>

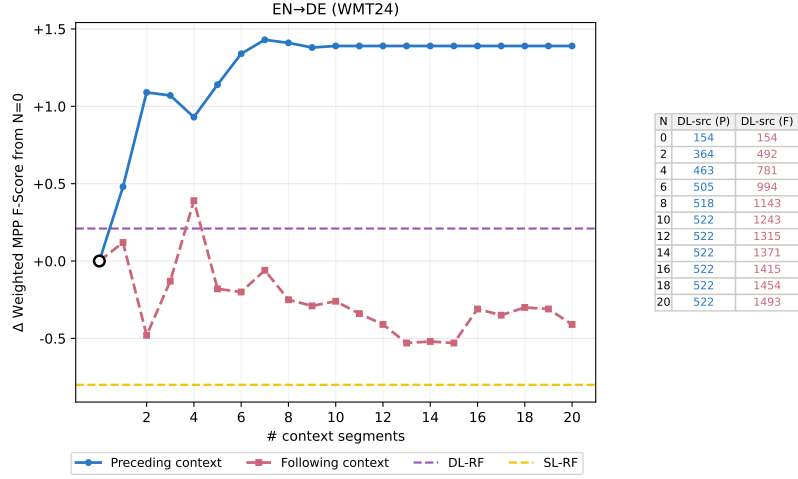
Table 12: Weighted micro-averaged MPP  $F$ -score on the WMT24 test set across different models and evaluation scenarios. The “Think” column indicates whether the model was run with thinking enabled. The best result for each language pair is shown in bold, while the best result for each model is underlined.

LP	Model	Think	Evaluation Scenarios			
			SL-RF	SL-RB	DL-RF	DL-RB
EN→KO	<i>Baseline</i>					
	xCOMET-XL	×	8.76	<u>10.03</u>	-	-
	<i>Qwen3</i>					
	Qwen3-8B	×	<u>14.26</u>	14.24	12.45	13.06
	Qwen3-8B	✓	12.50	<u>14.64</u>	10.52	11.71
	Qwen3-32B	×	17.36	<u>19.31</u>	15.32	17.22
	Qwen3-32B	✓	13.10	<u>16.67</u>	11.24	16.13
	<i>Gemma-4</i>					
	Gemma-4-E4B	×	16.39	<b>19.39</b>	11.69	12.54
	Gemma-4-E4B	✓	11.56	<u>15.83</u>	8.83	11.68
JA→ZH	<i>Qwen3 Fine-tuned</i>					
	Qwen3-8B-fine-tuned	×	13.58	14.24	14.28	<u>15.67</u>
	<i>Baseline</i>					
	xCOMET-XL	×	15.33	<u>18.73</u>	-	-
	<i>Qwen3</i>					
	Qwen3-8B	×	14.66	<u>22.62</u>	11.02	16.97
	Qwen3-8B	✓	15.92	<u>22.75</u>	11.20	16.50
	Qwen3-32B	×	22.64	<b>26.69</b>	21.32	23.69
	Qwen3-32B	✓	14.29	<u>23.80</u>	13.26	21.15
	<i>Gemma-4</i>					
	Gemma-4-E4B	×	17.37	<u>22.12</u>	11.00	13.11
	Gemma-4-E4B	✓	14.03	<u>19.51</u>	12.44	14.94
	<i>Qwen3 Fine-tuned</i>					
	Qwen3-8B-fine-tuned	×	18.82	<u>21.33</u>	19.30	20.65

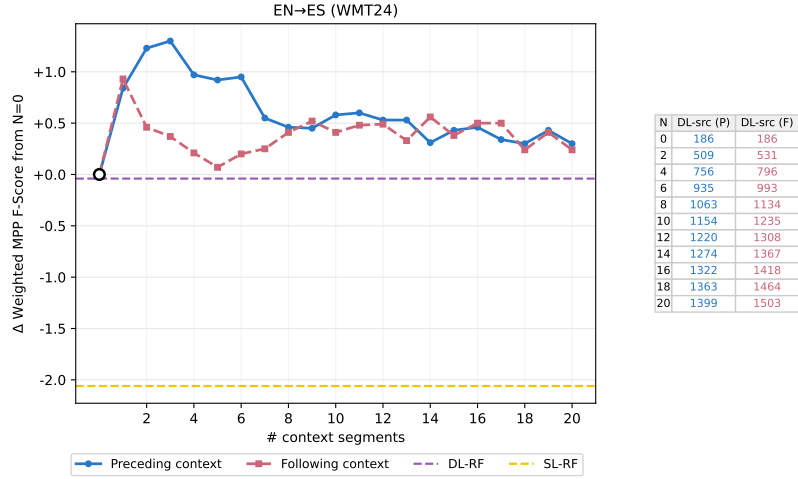
Table 13: Weighted micro-averaged MPP  $F$ -score on the WMT25 test set across different models and evaluation scenarios. The “Think” column indicates whether the model was run with thinking enabled. The best result for each language pair is shown in bold, while the best result for each model is underlined.

Test set	Effect	$\Delta$ [95% CI]
WMT24	DL – SL (overall)	+1.07 [0.65, 1.49]
	DL – SL   RF	+1.23 [0.74, 1.74]
	DL – SL   RB	+0.92 [0.51, 1.31]
	DL – SL   Non-Thinking	+0.64 [0.19, 1.09]
	DL – SL   Thinking	+1.50 [1.03, 1.98]
	DL – SL   Fine-tuned	+2.87 [2.22, 3.54]
WMT25	DL – SL (overall)	−3.44 [−3.96, −2.67]
	DL – SL   RF	−2.82 [−3.37, −1.97]
	DL – SL   RB	−4.07 [−4.68, −3.23]
	DL – SL   Non-Thinking	−3.97 [−4.56, −2.98]
	DL – SL   Thinking	−2.92 [−3.49, −2.26]
	DL – SL   Fine-tuned	+0.48 [−0.14, 1.13]

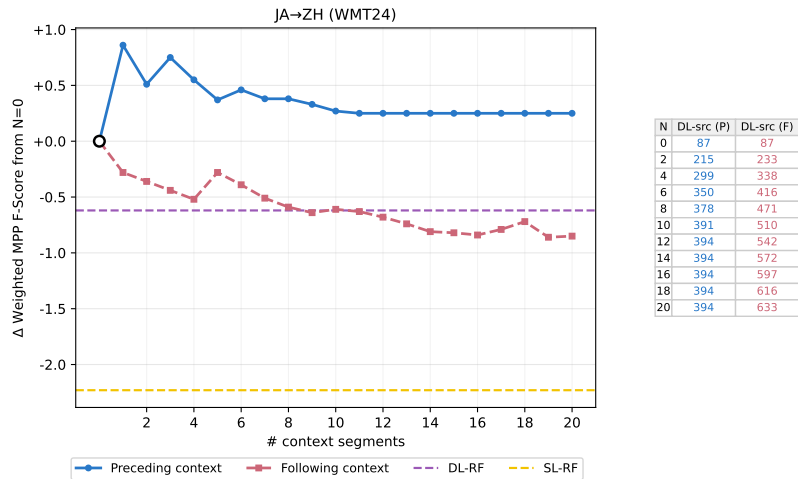
Table 14: Bootstrap estimates and 95% confidence intervals for the DL – SL (document-context) effect on weighted MPP  $F$ -score, under each conditioning scenario in Figure 1.



(a) EN→DE

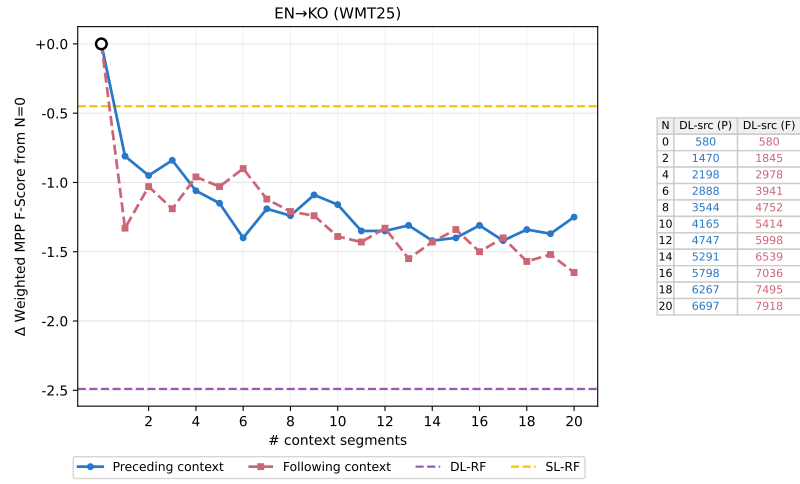


(b) EN→ES

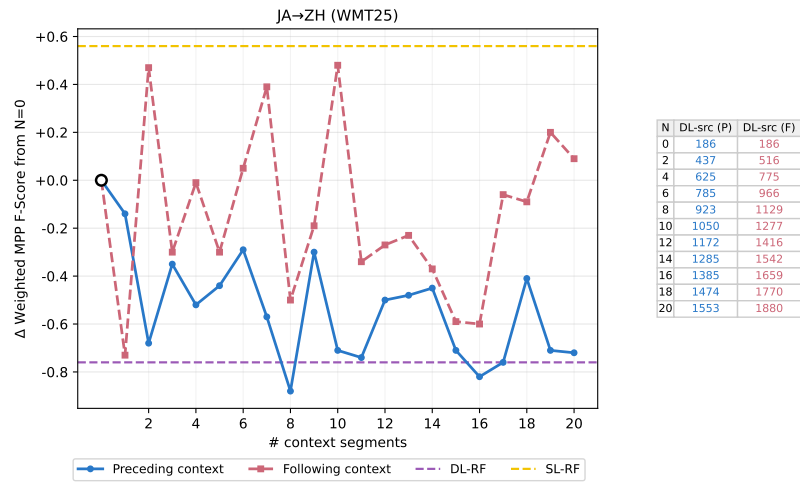


(c) JA→ZH

Figure 6: Effect of adding preceding and following context segments on weighted MPP  $F$ -score, relative to  $\text{SL-RF}_{\text{rep}}$ , across the WMT24 language pairs. "DL-src (P/F)" denotes the average source context length when using preceding (P) or following (F) context, respectively.



(a) EN→KO



(b) JA→ZH

Figure 7: Effect of adding preceding and following context segments on weighted MPP  $F$ -score, relative to  $\text{SL-RF}_{\text{rep}}$ , across the WMT25 language pairs. "DL-src (P/F)" denotes the average source context length when using preceding (P) or following (F) context, respectively.