Extending Automatic Machine Translation Evaluation to Book-Length Documents

Anonymous ACL submission

Abstract

Despite Large Language Models (LLMs) demonstrating superior translation performance and long-context capabilities, evaluation methodologies remain constrained to sentence-level assessment due to dataset limitations, token number restrictions in metrics, and rigid sentence boundary requirements. We introduce an evaluation scheme that extends existing automatic metrics to long-document translation by treating documents as continuous text and applying sentence segmentation and alignment methods. Our approach enables previously unattainable document-level evaluation, handling translations of arbitrary length generated with document-level prompts while accounting for under-/over-translations and varied sentence boundaries. Experiments show our scheme significantly outperforms existing long-form document evaluation schemes, while comparable to evaluations performed with groundtruth sentence alignments. Additionally, we apply our scheme to book-length texts and newly demonstrate that many open-weight LLMs fail to effectively translate documents at their reported maximum context lengths.

1 Introduction

Since the inception of Large Language Models (LLMs), the paradigm of machine translation (MT) has been shifting toward an LLM-based approach. In the WMT 2024 general translation shared task (Kocmi et al., 2024), LLM-based systems demonstrated strong performance, ultimately dominating submissions across all language pairs. Additionally, because of their long context windows, LLM-based systems may potentially be able to generate translations that better capture discourse-level phenomena and maintain coherence across longer spans of text. This development aligns with the long-standing trend in MT research to move beyond sentence-level processing toward *paragraph-level* (Deutsch et al., 2023), *discourse-level* (Bawden et al., 2018), and *document-level* (Zhu et al., 2024) translation.

However, despite claims that modern LLMs can process inputs of up to 1M tokens (Yang et al., 2025), evaluations of LLM-based translations remain largely confined to *sentence-level* or *segment-level*, in that they are only prompted to translate one sentence or segment at a time. This forms a significant gap between what LLMs can generate and what existing metrics can evaluate. This limitation stems from several challenges:

1. Many existing MT datasets (e.g., FLORES (Costa-jussà et al., 2022)) are inherently designed only for sentence-level assessment.
2. Commonly used model-based evaluation metrics have relatively low maximum token length limitations (e.g., 512 tokens for COMET).
3. Current automatic evaluation metrics require adherence to pre-defined sentence boundaries. This forces evaluators to either use sentence-level prompting or add artificial boundaries.

The first issue is partially addressed with the introduction of new datasets such as WMT24++ (Deutsch et al., 2025), which contains full documents instead of isolated sentences. However, significant lack of datasets remains when evaluating a model's capability to translate longer documents, such as book-length texts. Meanwhile, the issues of token length limitations and requirements for rigid sentence boundaries remain unresolved.

In this paper, we propose a solution to these challenges by introducing a scheme that extends existing automatic evaluation metrics to long documents. To summarize, our approach applies to arbitrarily long documents by using sentence segmenters and aligners to create appropriate sentence-level alignments. We treat those automatically aligned sentence pairs in two different ways. In the case where a valid sentence alignment is found, we apply the existing evaluation metric to the sentence

pair. In the case where translation errors occur - either when content from the source text is missing in the translation (*under-translation*) or when there is hallucinated content that is not present in the source (*over-translation*) - we detect these as null alignments and assign a fixed penalty. At the end, along with metric scores, we also report an auxiliary metric that reflect the ratio of null alignments to help track these over-translation and under-translation errors, which current MT evaluation metrics are having trouble detecting reliably.

Our experiments demonstrate that this scheme evaluates translations with comparable performance to existing sentence-level metrics when applied to cases with over-translation and under-translation. In addition, it handles cases where LLMs are liberal with sentence-boundaries, which create many-to-one and one-to-many sentence alignments. Lastly, we newly demonstrate that we can successfully apply this scheme to evaluate translations of book-length texts, and reveal that many open-weight LLMs cannot translate documents of their reported context length, because the number of under- and over-translation errors rise sharply as the input length gets longer. Our code and artifacts will be available at `anonymous.url`.

2 Related Work

2.1 Document-Level Translation

Document-level MT extends translation beyond isolated sentences by leveraging broader context for coherence. Existing approaches include simply concatenating adjacent sentences as a larger input to a standard MT model (Scherrer et al., 2019; Junczys-Dowmunt, 2019; Sun et al., 2022), as well as more advanced architectures that introduce context-specific modules: multi-encoder models encode previous sentences with separate encoders and hybrid attention mechanisms (Jean et al., 2017; Bawden et al., 2018; Voita et al., 2019; Miculicich et al., 2018; Maruf et al., 2019; Herold and Ney, 2023). Recent work has also focused on improving the quality of document-level translation by utilizing larger-scale document-level corpus (Thai et al., 2022; Al Ghussin et al., 2023; Post and Junczys-Dowmunt, 2023; Pal et al., 2024), as well as leveraging large language models (LLMs) (Karpinska and Iyyer, 2023; Wang et al., 2023).

Despite the progress, document-level translation still has a few limitations. First, a lot of work stick to a relatively small number of maximum input/output length. For example, Scherrer et al. (2019); Post and Junczys-Dowmunt (2023) both have maximum context length of 250 tokens, while Al Ghussin et al. (2023); Pal et al. (2024) have 512. Besides, some work (Junczys-Dowmunt, 2019; Post and Junczys-Dowmunt, 2023) introduce artificial sentence boundaries to the input, which provides native sentence segmentations for evaluation. This requires specialized training data or prompt, and there is no guarantee that the system will generate matching sentence boundaries as the input document.

2.2 Machine Translation Evaluation

Machine translation evaluation has shifted from string-based metrics (e.g., BLEU (Papineni et al., 2002), chrF (Popović, 2015)) to model-based metrics (e.g., COMET (Rei et al., 2020), MetricX (Juraska et al., 2024), GEMBA (Kocmi and Federmann, 2023)). Human evaluations like direct assessment (DA) and multi-dimensional quality metrics (MQM) played a crucial role in this paradigm shift by providing meta-evaluations and training data for model-based metrics.

Most model-based metrics are trained and evaluated on the segment-level. For example, COMET limits each input (source, target, and reference) to 512 tokens, while MetricX has a combined limit of 1,536 tokens across all inputs. In contrast, Qwen-2.5 (Yang et al., 2024), a recent open-source LLM, can handle input of 131,072 tokens and generate up to 8,192 tokens. Prior efforts have explored extending MT evaluation metrics beyond sentence-level. Vernikos et al. (2022) proposed adding prior sentences as context when training model-based metrics. Deutsch et al. (2023) trained metrics on paragraph-level data but found limited benefits. These studies are orthogonal to ours – they focus on building new model-based metrics with longer maximum input length, while we focus on applying existing metrics to long-form text.

Closest to the spirit of our work is MWERSegmenter (Matusov et al., 2005). It is a joint sentence segmentation and alignment scheme that has been the long-standing evaluation standard for unsegmented speech translation[1]. The high-level idea is to jointly segment and align long-form model output by minimizing the word error rate (WER) between the segmented text and the already-

[1]Specifically, MWERSegmenter has been the evaluation standard for the IWSLT speech translation shared tasks (`https://iwslt.org/`)

segmented reference text. The assumption behind the idea is that perfectly segmented and aligned sentences are more likely to be translated well, and thus should have a low WER. Similar to MWERSegmenter, Wang et al. (2023) implemented a segmentation and alignment scheme based on Bleualign (Sennrich and Volk, 2010), but there was no extensive discussion regarding the validity of the scheme. Apart from that, Raunak et al. (2024) proposed to extend existing metrics based on running evaluations on aligned sliding windows over sentences in a document, but the algorithm is still limited to the sentence-level prompting paradigm.

A few recent investigations (Salesky et al., 2023; Sperber et al., 2024) of MWERSegmenter in the context of long-form audio data raised concerns about the segmentation quality. The reader shall see that our results corroborate the concerns.

2.3 Long-Context LLM Evaluation

Recent progress in extending LLM architectures to handle longer contexts has spurred considerable research interest. Parallel to architectural advances, there has been growing attention toward systematically evaluating the capabilities of LLMs on long-context tasks. Kamradt (2023) developed an evaluation focusing on models' abilities to retrieve deeply nested information. Similarly, Bai et al. (2024) introduced a long-context bilingual benchmark for assessing models' comprehension and reasoning abilities, while An et al. (2024) shows that standardized evaluation criteria across multiple long-context scenarios are essential for comprehensive model assessment. Furthermore, Hsieh et al. (2024) highlights that there are discrepancies between theoretical capabilities and effective usable context lengths of contemporary LLMs.

Despite these advances, the evaluation methodologies have predominantly focused on general comprehension tasks rather than specialized applications like long-context machine translation. Existing metrics face limitations such as fixed maximum token lengths and rigid assumptions about sentence boundaries, which hinder effective evaluation of extensive, continuous texts, like books.

3 Preliminaries

Our ultimate goal is to find a way to evaluate the translation quality in the following scenarios:

- For translations of documents of arbitrary length generated with document-level prompts, while handling under-/over-translations and varied sentence boundaries
- For both reference-based and reference-free evaluations, thus enabling broader applications like curating high-quality training data for LLMs (similar to Finkelstein et al., 2024)

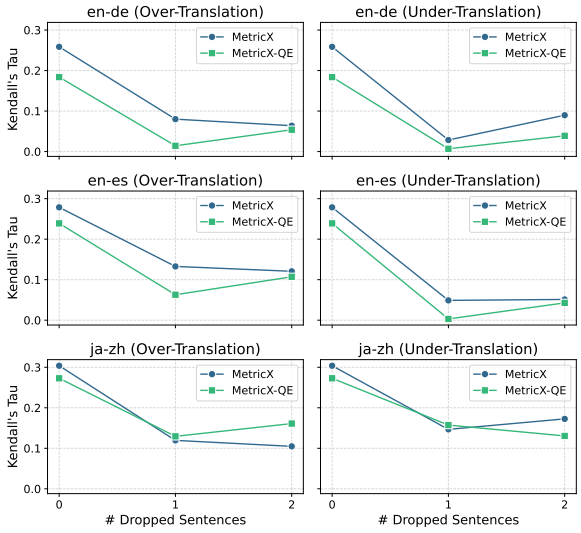


Figure 1: The Kendall's τ correlations of MetricX-24 and MetricX-24-QE show limited sensitivity to over- and under-translation; sentences with more than three drops are insufficient to estimate correlations reliably.

We start by justifying why extensions of existing metrics are required for document-level MT evaluation, rather than directly feeding concatenated sentence pairs from a document into existing metrics. A key limitation of the direct concatenation approach is that commonly used model-based evaluation metrics have relatively low maximum token length limits. Additionally, even for documents that are within the maximum token length limit, we show with a preliminary study in this section that directly applying state-of-the-art MT metrics to concatenated sentences is actually not able to reliably detect under- and over-translation errors.[2]

We evaluate the performance of MetricX-24 (Juraska et al., 2024) on such concatenation approach. To avoid going over the maximum token length limit of MetricX-24, we filter out cases where the concatenation of source and target inputs exceed 1024 tokens in length. We compute both MetricX-24 and its reference-free variant,

[2]The conclusion may seem different from a prior study Deutsch et al. (2023), but it's actually not a direct contradiction, because the evaluation in Deutsch et al. (2023) focuses only on cases where one-to-one mapping between source, target, and reference exists.

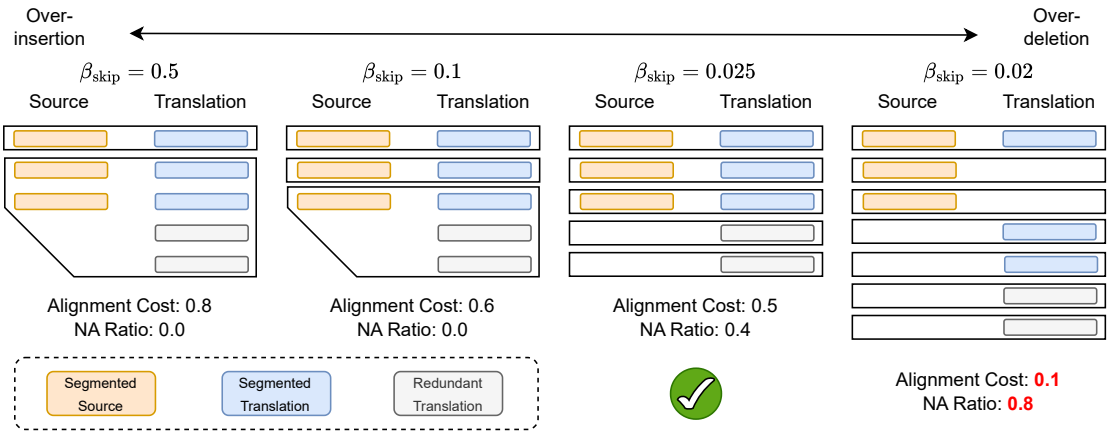


Figure 2: The effect of the skip cost (β_{skip}) on alignment behavior under over-translation. Higher skip costs increase the risk of over-insertion by allowing loose semantic matches to align, while lower skip costs enforce stricter alignment, leading to over-deletion. Over-deletion is indicated by spikes in the null alignment ratio (NA ratio) and low alignment costs, both shown in red.

MetricX-24-QE, across three language pairs: English–German (en-de), English–Spanish (en-es), and Japanese–Chinese (ja-zh). We use the dataset from the WMT 2024 Metrics Shared Task (Freitag et al., 2024). To simulate translation errors, we manipulate the texts in two ways: for over-translation, we remove one or two sentences from both source and reference texts, while for under-translation, we remove sentences only from the target side.[3] We measure the performance of the metrics by computing Kendall's τ correlations between the metric outputs and human evaluation scores.[4]

The results of this preliminary study (Figure 1) confirm that MetricX-24 and MetricX-24-QE have limited sensitivity to over- and under-translations, even within token length limits. This empirical evidence shows that the direct concatenation approach is inadequate for document-level MT evaluation. To apply existing MT metrics to document-level MT evaluation, we need an extension scheme that can properly handle these translation errors while working within the constraints of existing metrics.

4 Method

We now turn to our proposed extension scheme. Our approach consists of the following steps:

1. Segment the document-level system translation into indivudual sentences
2. Align the source and system translations
3. Apply existing metrics to the individual aligned sentences, then average sentence-level scores to obtain a document-level score

4.1 Sentence Segmentation

We use off-the-shelf sentence segmentation models to segment documents into sentences. We experimented with ersatz (Wicks and Post, 2021) and spaCy (Honnibal et al., 2020).

4.2 Sentence Alignment

Given a sentence-segmented source document $\mathbf{S} = \{s_1, \ldots, s_N\}$ and its translation $\mathbf{T} = \{t_1, \ldots, t_M\}$, the goal of sentence alignment is to identify a minimal-cost alignment path that maps contiguous spans of source sentences to contiguous spans of target sentences. We use Vecalign (Thompson and Koehn, 2019) to perform such alignment, with a few changes detailed below.

Adaptive Penalty Search Ideally, all over- and under-translations errors will result in *null alignments*. In Vecalign, null alignments are modeled via a skip cost, which is parameterized by a percentile-based threshold β_{skip}. If the skip cost is set too high, many alignments are forced between unrelated sentence blocks – essentially reverting to the scenario in our preliminary experiment. Conversely, if the skip cost becomes too low, the aligner will start to assign null alignments even to semantically related sentence pairs. Figure 2 illustrates an example of over-translation and demonstrates how different β_{skip} values impact the alignment result.

[3] We limit ourselves to removing at most two sentences since removing more would leave too few documents for reliable Kendall's τ correlation calculations.

[4] For more details on meta-evaluation, see Section 5.1.

Given that the optimal value of β_{skip} can vary depending on the severity of over- or under-translation in each individual document, we implement an adaptive search strategy to enhance robustness. We leverage the insight that over-deletion is often signaled by sudden spikes in the null alignment ratio and abnormally low average alignment costs. Since the optimal alignment typically occurs just before over-deletion sets in, our approach starts with a relatively high β_{skip} and progressively decreases it in small steps.

At each step, alignment quality is monitored using two heuristics to determine whether to terminate the search: (a) when the average alignment cost drops below a threshold, indicating excessive skipping, or (b) when the null alignment ratio exceeds a predefined limit at a step. Both patterns typically suggest that the skip penalty has become too lenient. In such cases, we revert to the previous step and treat it as the final alignment result. See Appendix B for implementation details and heuristic settings.

Building Better Text Embeddings Text embedding models are crucial for sentence alignment. We observe that sentence segmentation granularities vary across languages, which strains existing text embedding models. For example, suppose we have a long source sentence s that should align to smaller target sentences $\{t_1, \ldots, t_M\}$. In Vecalign, the scoring function calculates similarity between s and all consecutive blocks of $\{t_1, \ldots, t_M\}$. This is not what text embedding models are trained for, leading to suboptimal alignments.

Motivated by this observation, we build our own text embedding model that is specifically designed to handle the sentence segmentation granularities we described above. Our model is fine-tuned from BGE-M3 (Chen et al., 2024), which achieves high performance on bitext-mining task with only 568M parameters and without relying on instructions. The fine-tuning is performed on a synthetic dataset with query, positive and negative example triplets built from the News Commentary v18[5] dataset (see more details in Appendix B.3), with the FlagEmbedding toolkit[6]. The readers shall see that our fine-tuned text embedding model outperforms both LASER (Artetxe and Schwenk, 2019) and BGE-M3 text embedding models in our experiments.

[5] https://data.statmt.org/news-commentary/v18.1/

[6] https://github.com/FlagOpen/FlagEmbedding

4.3 Evaluation via Existing Metrics

Once the target translation is segmented and aligned with the source document, we calculate the segment-level translation quality using existing metrics, then average the scores to obtain a document-level score. Two numbers are reported per document: the average segment score and ratio of null alignments over aligned sentence pairs ("NA ratio"). For each null alignment, we assign the worst possible score (0 for COMET, 25 for MetricX) and include it in the average calculation.

5 Experiments

Our experiments are aimed to test if the proposed evaluation scheme can achieve the two goals stated in Section 3 – in other words, whether it is (1) robust to all kinds of anomalies in system translations and (2) effective with both reference-based and reference-free metrics. Our data and metric setup reflect the above goals.

To establish meaningful comparisons, we compare with two baselines. One is calculating metric scores using the groundtruth sentence boundaries and alignments provided by the dataset ("Gold"), which serves as a performance upper bound.[7] The other is calculating metric scores using the sentence boundaries and alignments derived by MWERSegmenter ("MWER").

Our experimental results demonstrate that our method consistently outperforms MWER while achieving comparable performance to Gold, validating its effectiveness. We further conduct an ablation study to examine how different sentence embeddings and segmenters affect performance.

5.1 Setup

Dataset We use the same dataset from preliminary experiments in Section 3. In all experiments, we merge existing sentence boundaries in the system translation to simulate system translations generated at document-level. We adhere to the same sentence boundaries on the source and reference sides during evaluation.

There are significant limitations if we only conduct meta-evaluation on the original test set, because the original test set is always guaranteed to have perfect sentence alignments (i.e., no null

[7] As the reader shall see, there are times when our score is higher than the upper bound performance. This is likely caused by the sentence segmentation variations between the sentence segmenter and boundaries in the test set. It shouldn't be interpreted as our method being better in a meaningful way.

	COMET	COMET-QE	MetricX	MetricX-QE	NA Ratio ($\lvert\Delta_{\text{Gold}}\rvert$)
Original					
Gold	0.3107	0.2785	0.3131	0.2748	0.0% (–)
Ours	0.3113	0.2773	0.3139	0.2728	0.7% (0.7%)
MWER	0.2964	0.2661	0.2965	0.2565	0.0% (0.0%)
Over-Translate					
Gold	0.3843	0.3551	0.3603	0.3037	10.0% (–)
Ours	0.3572	0.3409	0.3528	0.3141	11.2% (1.2%)
MWER	0.3501	0.3261	0.3209	0.2711	0.0% (10.0%)
Under-Translate					
Gold	0.3543	0.3216	0.3043	0.2857	10.0% (–)
Ours	0.3509	0.3347	0.3239	0.2861	5.8% (4.2%)
MWER	0.1893	0.1729	0.1463	0.1268	2.5% (7.5%)
Flex-Boundary					
Gold	0.3093	0.2778	0.3113	0.2726	0.0% (–)
Ours	0.3046	0.2728	0.3073	0.2670	1.2% (1.2%)
MWER	0.2589	0.2339	0.2551	0.2187	0.0% (0.0%)

Table 1: Correlation between document-level scores and human judgments under different evaluation settings. The first four columns are Kendall's τ correlation coefficients (↑), with the last column being the average NA ratio and the absolute difference from the groundtruth ($\lvert\Delta_{\text{Gold}}\rvert$). All numbers are averaged across three language pairs (en-de, en-es, ja-zh), and all reported numbers of our method are calculated with ersatz sentence segmenter and our fine-tuned BGE-M3 text embedding model.

alignments). Hence, in addition to the original test set ("original" case), we create three synthetic test sets by introducing anomalies into the original test set, namely "over-translate", "under-translate", and "flex-boundary" cases. The first two cases are created by randomly removing 10% of the sentences from the source/reference sides and system translations, respectively. The last case is created by merging 10% of the neighboring sentences in the system translations with GPT-4o. For more details, please refer to Appendix A.

Metric Our experiments cover both reference-based and reference-free ("QE") variants of COMET and MetricX.

Meta-Evaluation Similar to previous work and preliminary experiments, we use correlation between document-level scores and human judgments as the primary metric. Although both system translation and human judgments are performed at segment-level, previous work (Deutsch et al., 2023) has shown that MQM annotations are done with context of surrounding sentences, and sentences appear in document order. Hence, they are a good proxy for document translation quality. For cases with introduced null alignments, we assign 25 as the human-annotated MQM score for each null alignment, which is then converted into z-score in accordance with each human annotator's scoring distribution. Like previous work, we average the segment-level human judgment scores as the document-level scores.

We also report NA ratio for each method as the auxiliary metric. Ideally, we would like to achieve the same NA ratio as the groundtruth ($\lvert\Delta_{\text{Gold}}\rvert = 0$), but the reader should note the following caveats:

- Perfect NA ratio on its own doesn't necessarily imply a good evaluation scheme.[8] The correlation with human judgments is the ultimate measure of a good evaluation scheme.
- NA ratio is sometimes ill-defined for MWER.

5.2 Results

Main Results Table 1 shows a concise version of our main results (averaged across three language pairs). In terms of correlation with human judgments, our method achieves near-ideal performance, outperforming MWER while maintaining comparable correlation with human judgments to Gold. The trend is especially clear for "under-translate" and "flex-boundary" cases, where MWER suffers significant performance drops while our method remains robust. For a detailed version with per-language-pair breakdown, please refer to Appendix C. The readers shall see

[8] For example, in the "original" case, a very bad hypothetical evaluation scheme that aligns a random segment to the source can achieve the same 0% NA ratio as groundtruth.

	COMET	MetricX	NA Ratio ($\|\Delta_{\text{Gold}}\|$)
Original			
ersatz+LASER	0.3021	0.3067	1.2% (1.2%)
ersatz+BGE-m3	0.3021	0.3094	1.3% (1.3%)
ersatz+BGE-m3-ft	**0.3113**	**0.3139**	**0.7% (0.7%)**
spacy+BGE-m3-ft	0.3066	0.3096	1.4% (1.4%)
Over-Translate			
ersatz+LASER	0.3313	0.3370	8.6% (1.4%)
ersatz+BGE-m3	0.3212	0.3234	9.8% (0.2%)
ersatz+BGE-m3-ft	**0.3572**	**0.3528**	11.2% (1.2%)
spacy+BGE-m3-ft	0.3555	0.3508	**10.1% (0.1%)**
Under-Translate			
ersatz+LASER	0.3414	0.3145	**6.3% (3.7%)**
ersatz+BGE-m3	0.3347	0.3094	4.2% (5.8%)
ersatz+BGE-m3-ft	**0.3509**	**0.3239**	5.8% (4.2%)
spacy+BGE-m3-ft	0.3483	0.3215	6.0% (4.0%)
Flex-Boundary			
ersatz+LASER	0.3000	0.3043	1.9% (1.9%)
ersatz+BGE-m3	0.2979	0.3049	2.2% (2.2%)
ersatz+BGE-m3-ft	**0.3046**	**0.3073**	**1.2% (1.2%)**
spacy+BGE-m3-ft	0.3022	0.3050	1.5% (1.5%)

Table 2: Ablation study on different sentence embeddings and segmenters. Numbers are calculated similarly to Table 1 but only include reference-based metrics due to space limits. Boldface numbers indicate the highest correlation for the first two columns, and the NA ratio with the smallest $|\Delta_{\text{Gold}}|$ for the last column.

that the general trend shown in Table 1 is consistent across all language pairs.

For NA ratio, we can observe that MWER perfectly matches the groundtruth in two settings ("original" and "flex-boundary"). However, upon closer inspection, we conclude that this is not because MWER can more accurately estimate the NA ratio, but rather because MWER was not designed to account for certain translation anomalies. For example, MWER by design is not able to handle null alignments on the source side. Hence, in the "over-translate" case, MWERSegmenter simply merges over-translating system output to an arbitrary neighboring sentence, resulting in worse correlation with human judgments. Another such example lies in the "under-translate" case, where MWERSegmenter often segments a single system output into random small chunks. As for our method, while it is also not perfect with its NA ratio estimation, we argue that this auxiliary metric is still an useful indicator as to when under-/over-translation starts to get prevalent. Besides, compared to MWER, the misalignments introduced by our method are less likely to translate into catastrophic performance drops like MWER in the "under-translate" case.

While the lack of source code for MWERSegmenter makes it difficult to pinpoint the exact reason for its performance drop, looking at the segmented text, it is clear that MWER's algorithm struggles to distinguish between deletion and substitution errors, leading to erroneous choices in segmentation and alignment. Such case is worsened by poor translation quality, as the path to minimize WER becomes more obscure. This exemplifies the limitation of using WER, instead of a semantic-based score, as the scoring function for segmentation and alignment. Interestingly, while seemingly a symmetric case, insertion errors are less prone to this problem, likely because insertion errors are harder to distribute across multiple segments without introducing new errors.

Impact of Sentence Embedding Table 2 shows a comparison of our method with different sentence embeddings. It can be observed that our fine-tuned BGE-M3 embedding consistently outperforms LASER and the original BGE-M3 embedding in all data configurations. While the gap in correlation averaged across languages is small, we observe in the full breakdown (Table 4) that the effect of fine-tuning is most significant for ja-zh language pair. This happens to be the language pair with a lot of long sentences on the source side, which is more sensitive to the quality of sentence embedding. Being able to maintain robustness in that pair shows that our proposed embedding fine-tuning process successfully specializes the embedding model for the task of sentence alignment.

Impact of Sentence Segmenter Most of the numbers reported in this paper are calculated with the ersatz sentence segmenter. We also experimented with spaCy as the sentence segmenter as another ablation study, also shown in Table 2. We observed a small but consistent performance drop, likely due to the tendancy of spaCy segmenting sentences into smaller units, which does not align well with the long segments in WMT test sets.

6 Evaluation of Book-Length Translation Capability of Existing LLMs

Now that we have validated our evaluation method on WMT 2024 metrics shared task dataset, we briefly demonstrate that our method can be applied to assess the book-length translation capability of existing LLMs by conducting a similar experiment as Wang et al. (2024a). Our dataset comes from the Chinese-English (zh-en) section of the WMT 2024 Discourse-Level Literary Translation task (Wang et al., 2024b). Because the test set only contains

Model ID	Reference
utter-project/EuroLLM-9B-Instruct	Martins et al. (2024)
Qwen/Qwen2.5-14B-Instruct	Yang et al. (2024)
Qwen/Qwen2.5-72B-Instruct	Yang et al. (2024)
meta-llama/Llama-3.1-8B-Instruct	Dubey et al. (2024)
meta-llama/Llama-3.1-70B-Instruct	Dubey et al. (2024)
CohereForAI/aya-expanse-8b	Dang et al. (2024)
CohereForAI/aya-expanse-32b	Dang et al. (2024)

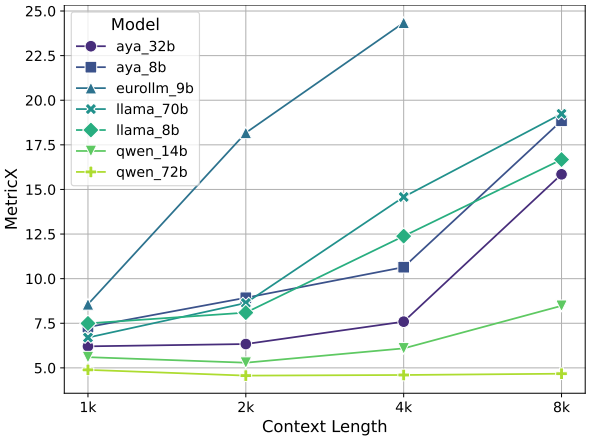
Table 3: List of LLMs evaluated for book-length translation capability.

book chapters instead of full books, we randomly pick a book with ID `2-xzltq` from the training split of the dataset and use it as our test set.

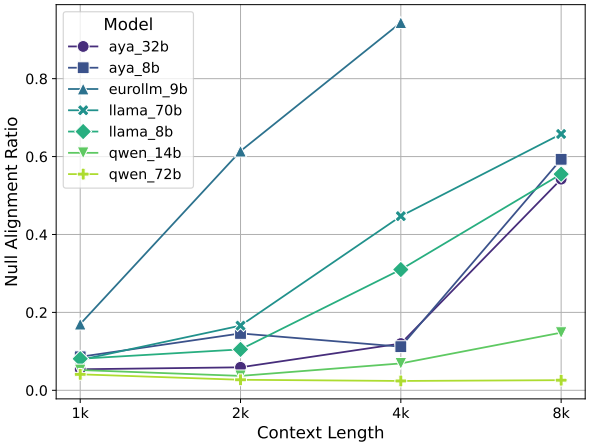
The LLMs evaluated are listed in Table 3. For simplicity, we adopted the same prompt and translation extraction procedure as used in WMT 2024 general machine translation shared task[9] for all the LLMs. Since current LLMs are constrained by maximum generation lengths and cannot translate the entire book in a single pass, we divide the content into segments of 1k, 2k, 4k, and 8k tokens, using tokenization from the tiktoken tokenizer[10]. Most of these models have a maximum generation length of 8k tokens, except for EuroLLM, which is capped at 4k.

Figure 3 shows the translation quality and NA ratio of the LLMs at different context lengths. Most models exhibit a sharp degradation in translation quality at context length of 4k or 8k. For example, at 4k context length, EuroLLM refuses to translate as instructed, but rather resorting to summarizing the input document in the target language. Comparing with the trend in NA ratio, it is also clear that under-translation/over-translation errors played a significant role in such degradation. The only noteworthy exception is `Qwen2.5-72B-Instruct`, which shows a much more stable performance across different context lengths. In fact, with the increasing context length up to 4k, there is a small improvement in translation quality, which shows its ability to utilize long-context information to obtain better translations.

This benchmark shows a significant gap between claimed max generation length and the actual capability of LLMs to translate long-context documents. As future work keeps improving LLM's long-context processing capabilities, we call on the



(a) Translation Quality (MetricX ↓)



(b) Null Alignment Ratio (↓)

Figure 3: LLM Translation Performance at Different Context Lengths

community to adopt this evaluation practice to gain better insights into such capabilities in downstream applications such as machine translation.

7 Conclusion

We propose a novel extension scheme that enables evaluation metrics to evaluate unsegmented document-level translations of arbitrary lengths. Our scheme works with any existing evaluation metric and eliminates the dependency on sentence-level prompting and pre-segmented reference translations. Experimental results show that our extension scheme achieves strong correlation with human judgments while demonstrating robustness to common LLM translation anomalies like over- and under-translation. Through this work, we aim to facilitate machine translation research in its ongoing shift away from sentence-level paradigm, while also offering new perspectives for evaluating LLMs' long-context generation capabilities.

[9] https://github.com/wmt-conference/wmt-collect-translations/blob/704b3825730f93a3ee3a0fda44af9414937b6d5a/tools/prompts.py#L23

[10] https://github.com/openai/tiktoken

Limitations

We acknowledge that an LLM-based metric based on long-context, open-source LLMs is a promising (and probably the eventual) solution to the problem of long-context MT evaluation. While previous work has shown that LLM-based metrics such as GEMBA (Kocmi and Federmann, 2023) or AutoMQM (Fernandes et al., 2023) can perform on-par as state-of-the-art BERT-based metrics such as COMET, they have to rely on GPT-4 or GPT-4o as the underlying LLM and are currently prohibitively expensive for MT evaluation of book-length documents. Their open-source LLM counterparts, on the other hand, are not able to match the performance of state-of-the-art metrics. Hence, we leave exploration of this direction as future work and focus on extending existing metrics for long-context MT evaluation in this study.

Our meta-evaluation is also limited in the sense that none of the metrics evaluated with our proposed extension scheme explicitly captures any discourse-level information (e.g. co-reference, consistency in word choice, etc.). This is partially due to the fact that most of the metrics that incorporate these information are targeted evaluations that depend on specific datasets to operate, and not easily extendable to WMT test sets. Another aspect worth considering is that the human judgments used in our evaluation do not explicitly instruct the annotators to consider discourse-level information. Hence, whether those extra information will show benefits in meta-evaluations based on current human judgments remains unclear.

The extra sentence segmentation and alignment step involved in our proposed extension scheme may introduce certain biases and noises. For example, since merging sentences during translation may result in alignment errors, the proposed extension scheme may unfairly prefer systems that adheres to the sentence boundaries in the original source document. We have extensively validated in our empirical experiments that such biases, if existing, would have minimal impact on the evaluation. We would also like to point out that even state-of-the-art MT metrics have their own biases (as pointed out in Section 3). Since existing work (Amrhein and Sennrich, 2022) has already proposed methods to reveal the pathologies of model-based MT metrics, we believe conducting a similar study with our proposed extension scheme will provide more comprehensive insights into its limitations.

The fine-tuned BGE-m3 embedding model used in our proposed extension scheme is largely a proof-of-concept, due to the fact that it is trained on a small dataset, covering only languages of our interest. We believe that a specialized text embedding model for sentence alignment is not only useful for our proposed extension scheme, but also for its more traditional use cases, such as curation of web-crawled data. In the future, we plan to explore extending the volume of the training data and supported languages to improve the usability of our proposed extension scheme.

References

Yusser Al Ghussin, Jingyi Zhang, and Josef van Genabith. 2023. Exploring paracrawl for document-level neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1304–1310, Dubrovnik, Croatia. Association for Computational Linguistics.

Chantal Amrhein and Rico Sennrich. 2022. Identifying weaknesses in machine translation metrics through minimum Bayes risk decoding: A case study for COMET. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1125–1141, Online only. Association for Computational Linguistics.

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2024. L-eval: Instituting standardized evaluation for long context language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14388–14411, Bangkok, Thailand. Association for Computational Linguistics.

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

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3119–3137, Bangkok, Thailand. 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.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Y. Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, and 19 others. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *CoRR*, abs/2412.04261.

Daniel Deutsch, Eleftheria Briakou, Isaac Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, and 1 others. 2025. Wmt24++: Expanding the language coverage of wmt24 to 55 languages & dialects. *arXiv preprint arXiv:2502.12404*.

Daniel Deutsch, Juraj Juraska, Mara Finkelstein, and Markus Freitag. 2023. Training and meta-evaluating machine translation evaluation metrics at the paragraph level. In *Proceedings of the Eighth Conference on Machine Translation*, pages 996–1013, Singapore. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 82 others. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

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

Mara Finkelstein, David Vilar, and Markus Freitag. 2024. Introducing the NewsPaLM MBR and QE dataset: LLM-generated high-quality parallel data outperforms traditional web-crawled data. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1355–1372, Miami, Florida, USA. Association for Computational Linguistics.

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.

Christian Herold and Hermann Ney. 2023. Improving long context document-level machine translation. In *Proceedings of the 4th Workshop on Computational Approaches to Discourse (CODI 2023)*, pages 112–125, Toronto, Canada. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. RULER: what's the real context size of your long-context language models? *CoRR*, abs/2404.06654.

Sébastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. Does neural machine translation benefit from larger context? *CoRR*, abs/1704.05135.

Marcin Junczys-Dowmunt. 2019. Microsoft translator at WMT 2019: Towards large-scale document-level neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 225–233, Florence, Italy. 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.

Gregory Kamradt. 2023. Needle in a haystack - pressure testing llms. *Github*.

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. In *Proceedings of the Eighth Conference on Machine Translation*, pages 419–451, Singapore. 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. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

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

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno Miguel Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, M. Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2024. Eurollm: Multilingual language models for europe. *CoRR*, abs/2409.16235.

Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. 2019. Selective attention for context-aware neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3092–3102, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954, Brussels, Belgium. Association for Computational Linguistics.

Proyag Pal, Alexandra Birch, and Kenneth Heafield. 2024. Document-level machine translation with large-scale public parallel corpora. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13185–13197, Bangkok, Thailand. Association for Computational Linguistics.

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

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post and Marcin Junczys-Dowmunt. 2023. Escaping the sentence-level paradigm in machine translation. *CoRR*, abs/2304.12959.

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, 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.

Elizabeth Salesky, Kareem Darwish, Mohamed Al-Badrashiny, Mona Diab, and Jan Niehues. 2023. Evaluating multilingual speech translation under realistic conditions with resegmentation and terminology. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 62–78, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Yves Scherrer, Jörg Tiedemann, and Sharid Loáiciga. 2019. Analysing concatenation approaches to document-level NMT in two different domains. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 51–61, Hong Kong, China. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of ACL*.

Rico Sennrich and Martin Volk. 2010. Mt-based sentence alignment for ocr-generated parallel texts. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas: Research Papers, AMTA 2010, Denver, Colorado, USA, October 31 - November 4, 2010*. Association for Machine Translation in the Americas.

Matthias Sperber, Ondřej Bojar, Barry Haddow, Dávid Javorský, Xutai Ma, Matteo Negri, Jan Niehues, Peter Polák, Elizabeth Salesky, Katsuhito Sudoh, and Marco Turchi. 2024. Evaluating the IWSLT2023 speech translation tasks: Human annotations, automatic metrics, and segmentation. In *Proceedings of*

the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 6484–6495, Torino, Italia. ELRA and ICCL.

Zewei Sun, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Lei Li. 2022. Rethinking document-level neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3537–3548, Dublin, Ireland. Association for Computational Linguistics.

Katherine Thai, Marzena Karpinska, Kalpesh Krishna, Bill Ray, Moira Inghilleri, John Wieting, and Mohit Iyyer. 2022. Exploring document-level literary machine translation with parallel paragraphs from world literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9882–9902, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.

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.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

Longyue Wang, Zefeng Du, Wenxiang Jiao, Chenyang Lyu, Jianhui Pang, Leyang Cui, Kaiqiang Song, Derek Wong, Shuming Shi, and Zhaopeng Tu. 2024a. Benchmarking and improving long-text translation with large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7175–7187, Bangkok, Thailand. Association for Computational Linguistics.

Longyue Wang, Siyou Liu, Chenyang Lyu, Wenxiang Jiao, Xing Wang, Jiahao Xu, Zhaopeng Tu, Yan Gu, Weiyu Chen, Minghao Wu, Liting Zhou, Philipp Koehn, Andy Way, and Yulin Yuan. 2024b. Findings of the WMT 2024 shared task on discourse-level literary translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 699–700, Miami, Florida, USA. Association for Computational Linguistics.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore. Association for Computational Linguistics.

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

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, Junyang Lin, Kai Dang, Kexin Yang, Le Yu, Mei Li, Minmin Sun, Qin Zhu, Rui Men, Tao He, and 9 others. 2025. Qwen2.5-1m technical report. *CoRR*, abs/2501.15383.

Dawei Zhu, Sony Trenous, Xiaoyu Shen, Dietrich Klakow, Bill Byrne, and Eva Hasler. 2024. A preference-driven paradigm for enhanced translation with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3385–3403, Mexico City, Mexico. Association for Computational Linguistics.

A Synthetic Test Data Creation

To evaluate the robustness of our evaluation framework, we construct synthetic test data simulating three common alignment challenges: under-/over-translation and varied sentence boundaries.

A.1 Synthetic Under- and Over-Translations

We simulate under- and over-translation scenarios by randomly removing 10% of the segments from either the source/reference sides or the system translations, respectively. To maintain meaningful context, we avoid sampling from documents that contain only a single segment.

A.2 Synthetic Sentence Boundary Variation

To simulate sentence boundary variation, we generate synthetic test data by merging 10% of adjacent segments in the source side. This process is conducted using GPT-4o, with the following constraints:

- Only segments that are neither the first nor the last in a document are eligible for merging.
- For each eligible candidate, an attempt is made to merge it with its subsequent segment.
- Merging is only accepted if the semantic difference between the merged and original segments is minimal. We use BLEURT (Sellam et al., 2020) to assess the semantic similarity, accepting only merges with a BLEURT score greater than 0.85. If a candidate pair fails this criterion, another eligible pair is sampled.

GPT-4o is instructed to merge adjacent segments into a single, fluent sentence without changing the original meaning, vocabulary, or the order of information. Figure 4 shows the prompt template used to guide the model.

Figure 5 provides an example of the merging process before and after GPT-4o rewriting. The sentences initially presented separately are transformed into a single sentence using appropriate transitional phrases.

This procedure enables us to test our evaluation method's robustness in conditions reflecting realistic variations in sentence boundary alignments while ensuring that human annotations remain valid and can be directly reused without recalibration.

B Implementation Details

In this section, we provide detailed implementation information on extending existing evaluation metrics to book-length documents. Our proposed approach is designed as a flexible framework where the underlying models can be readily substituted. Here, we specifically outline the experimental configurations used in this study.

B.1 Sentence Segmentation

For sentence segmentation, our experiments employ two different models: `SpaCy` and `ersatz`. `SpaCy` requires specification of the target language, whereas `ersatz` is language-agnostic, making it suitable for multilingual segmentation tasks.

The experiments in this paper cover five languages: English, German, Spanish, Japanese, and Chinese. Corresponding `SpaCy` models for each language are as follows:

- English (en): `en_core_web_sm`
- German (de): `de_core_news_sm`
- Spanish (es): `es_core_news_sm`
- Japanese (ja): `ja_core_news_sm`
- Chinese (zh): `zh_core_web_sm`

B.2 Sentence Alignment

We adopt a robust sentence alignment strategy based on Vecalign (Thompson and Koehn, 2019), leveraging multilingual sentence embeddings and an efficient dynamic programming approximation to identify many-to-many alignments between sentence segments.

In our experiments, we set the maximum number of allowed overlap size to 16. This allows us to search for source-target sentence alignments of size N–M, where $N + M \leq 16$. While we use a fixed overlap size in our experiments, it can be estimated from reference documents. Detailed explanations are provided in Appendix B.4.

Adaptive Penalty Search. In Vecalign, null alignments are handled via a skip cost, parameterized by a percentile-based threshold β_{skip}, representing a quantile in the empirical distribution of 1-1 alignment costs. To identify the optimal β_{skip}, we perform a search starting from 0.2, which is the default value in Vecalign, and progressively decrease it in small steps of 0.005. At each step, we detect signals of over-deletion. Upon detection, we revert to the previous step and treat it as the final alignment result.

System:
You are a helpful assistant.

User :
Please merge these two segments into one sentence while preserving their original meaning, word choice, and order.
Instead of simply concatenating them, use appropriate transitional expressions so that the segments are naturally connected without merely inserting a period or extra whitespace.
Ensure the final result flows coherently and no important information is omitted.
Return only the merged text on a single line, with no additional commentary or extraneous text.
First segment: <first segment>
Second segment: <second segment>

Figure 4: Prompt used for instructing GPT-4o to merge sentences.

First segment:
Move will also help transform land at the derelict Gartshore Works site.

Second segment:
A Scottish recycling business that has already processed more than a million tonnes of construction waste has opened a second plant following a multi-million pound investment.

After GPT-4o merging:
Move will also help transform land at the derelict Gartshore Works site **as** a Scottish recycling business that has already processed more than a million tonnes of construction waste has opened a second plant following a multi-million pound investment.

Figure 5: Example of merged sentences before and after GPT-4o rewriting.

Heuristic Termination Conditions. The search is terminated based on two heuristic signals that indicate over-deletion:

(a) The average alignment cost falls below 0.3.

(b) The null alignment ratio at a step exceeds 0.15.

Both patterns suggest that the skip cost has become too lenient, leading to excessive null alignments. In addition, two rare edge cases are handled with early stopping:

(a) If the average alignment cost increases rather than decreases at a step, indicating misalignments with semantically distant content.

(b) If the average alignment cost exceeds 0.7, indicating poor alignment quality overall.

Parameter Tuning. These heuristic termination conditions were tuned empirically on the test set from the WMT24 Discourse-Level Literary Translation shared task, using only the Chinese→English portion. To remain within the context length of our selected LLMs, we segment each instance from the training and validation sets into chunks of up to 1024 tokens. Translations are generated using `meta-llama/Llama-3.1-8B-Instruct` and `GPT-4o 2024-08-06`, and sentence embeddings are computed using LASER. We empirically validate alignment quality by manually inspecting whether translation errors are consistently marked as null alignments. Once verified, the same configuration is used for all experiments in this paper.

While the heuristic termination conditions are robust in our experiments, they may vary depending on the translation direction and source/reference sentence boundaries. These parameters can be further refined using reference translations, which are typically assumed to be perfectly aligned – i.e., with a null alignment (NA) ratio of zero. This provides a basis for estimating how much the NA ratio increases when the skip cost becomes too lenient, as well as the expected alignment cost under ideal conditions. These estimates can then inform the selection of appropriate heuristic parameters when evaluating future system translations.

B.3 Details for Text Embedding Model Fine-Tuning

We used News Commentary 18.1 parallel data from any language pairs that is a combination of Chinese (zh), English (en), German (de), Japanese (ja), and Spanish (es), which are the languages of interest in our evaluation. For each language pair, we use either the full dataset or only first 10,000 lines of the dataset, whichever is smaller. We build the example triplets with the following: we concatenate each of the two neighboring sentence pairs in the parallel corpus and use the source/target side as the query/positive example, respectively. As for the negative example, we always construct two vari-

ants: (1) we randomly drop on of the two sentences on the target side (2) we retrieve a nearby sentence by randomly looking forward 1 to 3 sentences in the dataset and use it to substitute the second sentence on the target side. The intuition behind these example triplets is for the model to better distinguish a good translation from (1) an incomplete translation, and (2) a sentence that has a similar topic but is not a translation. The resulting synthetic dataset has 130,436 examples.

We fine-tune the BGE-M3 embedding on the synthetic dataset with InfoNCE loss (van den Oord et al., 2018) for two epochs. We did not conduct an extensive hyperparameter search, but simply use the setup in the fine-tuning tutorial in the FlagEmbedding package[11]. We directly used the checkpoint at the end of the training without using a validation set to select the best checkpoint.

B.4 Setting Overlap Size for Alignment

Beyond the text embedding model, we also observe that the choice of overlap size in Vecalign can significantly impact the alignment quality. The overlap size defines the size of the blocks that are compared to each other in the alignment search. A small overlap size causes some ideal alignments to fall out of search space. For example, if the overlap size is set to 8, but a long source sentence should be aligned to a sentence block of size 16 on the target side, such groundtruth alignment will never be considered by the search algorithm. On the other hand, a large overlap size will increase the computational cost.

With datasets that comes with human-segmented sentence boundaries and alignment (which covers the vast majority of use cases), one can easily and accurately estimate the appropriate overlap size by re-segmenting the reference documents with sentence segmenter, calculate the maximum ratio of sentences as segmented by human and by the sentence segmenter. With datasets that does not come with human-segmented sentence boundaries or references, one would have to first conduct a pilot study with sentence-level translation to estimate the appropriate overlap size. However, the good news is that with a given sentence segmenter, the overlap size only needs to be estimated once per language, and can be reused for different datasets. Besides, in the case where estimation is not very accurate, one can always err on the safe side and set a larger overlap size.

[11] https://github.com/FlagOpen/FlagEmbedding/blob/024e789d599eb4cf9a208e98d27508ad455f5ecb/Tutorials/7_Fine-tuning/7.1.2_Fine-tune.ipynb

C Supplementary Results

Since the results in the main paper are condensed versions with averaging across different language pairs or showing only a subset of the metrics evaluated, we attach the full breakdown of the results in Table 4 for readers' reference.

D Licensing of Artifacts

Almost all code, model, and data artifacts we used in this paper are publicly available with permissive licenses (MIT/Apache 2.0/CC-BY-4.0). The only exceptions are Aya models (CC-BY-NC 4.0), and GPT-4o (OpenAI API Terms of Use), which still allows research use. We also plan to release all created code, model, and data artifacts under a permissive license.

E Use of AI Assistants

We used a code editor with generative AI functionalities during code development and paper writing (in the latter case, it only assists with LaTeX code completion and minor text editing). We also used various AI assistants for creating miscellaneous single-use data processing scripts, as well as all the figures in this paper. All AI-generated artifacts were carefully reviewed and accepted by the authors.

	COMET			COMET-QE			METRICX			METRICX-QE			NA Ratio		
	en-de	**en-es**	**ja-zh**	**en-de**	**en-es**	**ja-zh**	**en-de**	**en-es**	**ja-zh**	**en-de**	**en-es**	**ja-zh**	**en-de**	**en-es**	**ja-zh**
Original															
Gold	0.3188	0.2320	0.3813	0.2628	0.2248	0.3479	0.3487	0.2352	0.3553	0.2927	0.1995	0.3321	0%	0%	0%
MWER	0.2923	0.2376	0.3593	0.2349	0.2228	0.3406	0.3201	0.2413	0.3281	0.2632	0.2050	0.3012	0%	0%	0%
Ours-ersatz-BGE-M3-ft	0.3178	0.2410	0.3750	0.2612	0.2291	0.3417	0.3468	0.2454	0.3495	0.2896	0.2050	0.3237	0.25%	0.75%	1.32%
Ours-spacy-BGE-M3-ft	0.3128	0.2350	0.3719	0.2541	0.2261	0.3399	0.3417	0.2401	0.3469	0.2851	0.2035	0.3218	0.99%	0.81%	2.34%
Our-ersatz-LASER	0.3127	0.2340	0.3596	0.2555	0.2258	0.3340	0.3434	0.2392	0.3376	0.2887	0.2029	0.3133	0.23%	0.82%	2.40%
Ours-ersatz-BGE-M3	0.3141	0.2300	0.3623	0.2588	0.2205	0.3335	0.3465	0.2339	0.3478	0.2908	0.1977	0.3200	0.36%	0.81%	2.84%
Over-Translate															
Gold	0.3502	0.3975	0.4053	0.3110	0.3962	0.3580	0.3739	0.3430	0.3640	0.3094	0.2620	0.3398	10%	10%	10%
MWER	0.2857	0.3816	0.3831	0.2498	0.3747	0.3539	0.2873	0.3533	0.3222	0.2258	0.2957	0.2918	10%	10%	10%
Ours-ersatz-BGE-M3-ft	0.3353	0.3488	0.3874	0.3067	0.3628	0.3532	0.3635	0.3400	0.3549	0.3096	0.3003	0.3325	9.29%	13.28%	11.09%
Ours-spacy-BGE-M3-ft	0.3324	0.3495	0.3847	0.2978	0.3625	0.3504	0.3596	0.3405	0.3522	0.3058	0.3005	0.3300	9.47%	9.55%	11.35%
Our-ersatz-LASER	0.3161	0.3146	0.3633	0.2899	0.3361	0.3210	0.3496	0.3212	0.3403	0.2953	0.2752	0.3181	8.16%	9.07%	8.66%
Ours-ersatz-BGE-M3	0.2958	0.3102	0.3577	0.2668	0.3278	0.3199	0.3300	0.303	0.3372	0.2801	0.2574	0.3137	9.26%	8.61%	11.37%
Under-Translate															
Gold	0.3494	0.3349	0.3785	0.3227	0.3416	0.3004	0.3493	0.2564	0.3071	0.2946	0.2406	0.3220	10%	10%	10%
MWER	0.1728	0.1935	0.2016	0.1550	0.1826	0.1812	0.1619	0.1266	0.1503	0.1188	0.0950	0.1667	2.65%	3.64%	1.35%
Ours-ersatz-BGE-M3-ft	0.3334	0.3499	0.3693	0.3119	0.3619	0.3304	0.3427	0.2944	0.3347	0.2912	0.2566	0.3105	5.10%	6.62%	5.72%
Ours-spacy-BGE-M3-ft	0.3301	0.3481	0.3666	0.3058	0.3610	0.3288	0.3396	0.2923	0.3325	0.2871	0.2556	0.3107	5.47%	6.61%	6.01%
Our-ersatz-LASER	0.3312	0.3415	0.3516	0.3049	0.3524	0.3192	0.3359	0.2850	0.3227	0.2879	0.2529	0.2993	5.55%	6.34%	7.13%
Ours-ersatz-BGE-M3	0.3230	0.3225	0.3586	0.3041	0.3441	0.3167	0.3275	0.2726	0.3280	0.2829	0.2447	0.3050	2.98%	4.14%	5.54%
Flex-Boundary															
Gold	0.3148	0.2321	0.3809	0.2580	0.2270	0.3484	0.3458	0.2372	0.3510	0.2876	0.2011	0.3291	0%	0%	0%
MWER	0.2327	0.1853	0.3587	0.1788	0.1831	0.3399	0.2490	0.1929	0.3235	0.1930	0.1652	0.2978	0%	0%	0%
Ours-ersatz-BGE-M3-ft	0.3114	0.2299	0.3725	0.2546	0.2227	0.3411	0.3416	0.2379	0.3424	0.2841	0.2010	0.3160	0.98%	1.59%	1.05%
Ours-spacy-BGE-M3-ft	0.3062	0.2313	0.3690	0.2478	0.2237	0.3397	0.3359	0.2391	0.3399	0.2793	0.2022	0.3141	1.43%	1.64%	1.35%
Our-ersatz-LASER	0.3093	0.2300	0.3606	0.2537	0.2232	0.3342	0.3419	0.2373	0.3338	0.2868	0.2003	0.3093	1.31%	1.81%	2.55%
Ours-ersatz-BGE-M3	0.3079	0.2280	0.3577	0.2533	0.2183	0.3330	0.3414	0.2339	0.3393	0.2846	0.1971	0.3123	1.19%	2.03%	3.33%

Table 4: Full breakdown of our results on the WMT 2024 Metrics Shared Task dataset.