Align-then-Slide: A complete evaluation framework for Ultra-Long Document-Level Machine Translation

Anonymous ACL submission

Abstract

Large language models (LLMs) have ushered in a new era for document-level machine translation (*doc*-mt), yet their whole-document outputs challenge existing evaluation methods that assume sentence-by-sentence alignment. We introduce ***Align-then-Slide***, a complete evaluation framework for ultra-long *doc*-mt. In the Align stage, we automatically infer sentence-level source–target correspondences and rebuild the target to match the source sentence number, resolving omissions and many-to-one/one-to-many mappings. In the n-Chunk Sliding Evaluate stage, we calculate averaged metric scores under 1-, 2-, 3- and 4-chunk for multi-granularity assessment. On WMT benchmarks our rankings achieve a Pearson correlation of 0.929 with expert MQM scores; on a newly curated real-world test set they again align closely with human judgments. Notably, our method attained SOTA results in all 16 language directions of the segment-level quality-prediction track at WMT2025. When used directly as a reward model for GRPO, it yields translations preferred over a vanilla SFT baseline. These results validate Align-then-Slide as an accurate, robust and actionable evaluation tool for *doc*-mt systems.

1 Introduction

Large language models (LLMs) are opening a new chapter for document-level machine translation (*doc*-mt) (Kim et al., 2019; Maruf et al., 2022; Fernandes et al., 2021). Leveraging their exceptional capacity for long-context modeling and deep semantic understanding, LLMs can generate entire translations that are not only fluent and coherent but also faithful to the document's global meaning, far surpassing the limitations of conventional sentence-by-sentence approaches.

In document-level machine translation evaluation, the prevailing paradigm is to lift proven sentence-level metrics such as BLEU (Papineni et al., 2002), BERTScore (Zhang et al., 2019) and COMET (Rei et al., 2020a, 2022) to the document level. The latest studies often use LLMs for scoring (Gu et al., 2025), but this method has biases, inaccuracies, and low efficiency (GUO et al., 2025). Vernikos et al.'s *doc*-metrics achieves this with disarming simplicity: it merely prepends the preceding reference sentences to the current hypothesis–reference pair before encoding, instantly injecting document-wide context. Raunak et al.'s SLIDE adopts a chunk-wise strategy: a sliding window sweeps across the document, feeding contiguous blocks of sentences into an off-the-shelf quality-estimation model without any architectural tweaks, thereby enabling end-to-end document-level assessment.

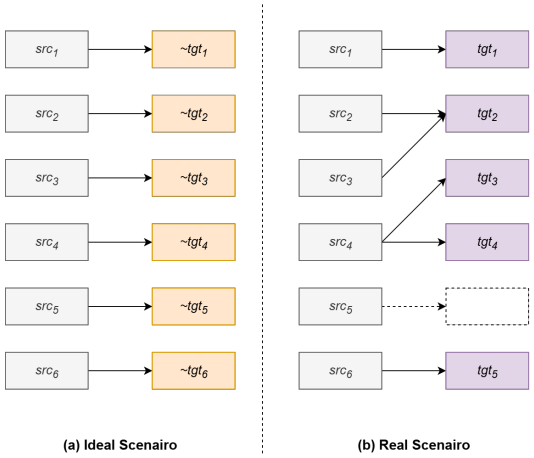

Figure 1: Ideal vs. Real scenarios between source and translated texts. (a) Perfect one-to-one correspondence assumed by prior metrics. (b) Actual complexities: whole-sentence omissions, many-to-one and one-to-many mappings, and variable target sentence counts.

Yet prior work implicitly assumes that the document has been segmented into sentences and that source and target sentences align one-to-one. Figure 1(a) depicts this Ideal Scenario. The Real Scenario, shown in Figure 1(b), introduces three thorny

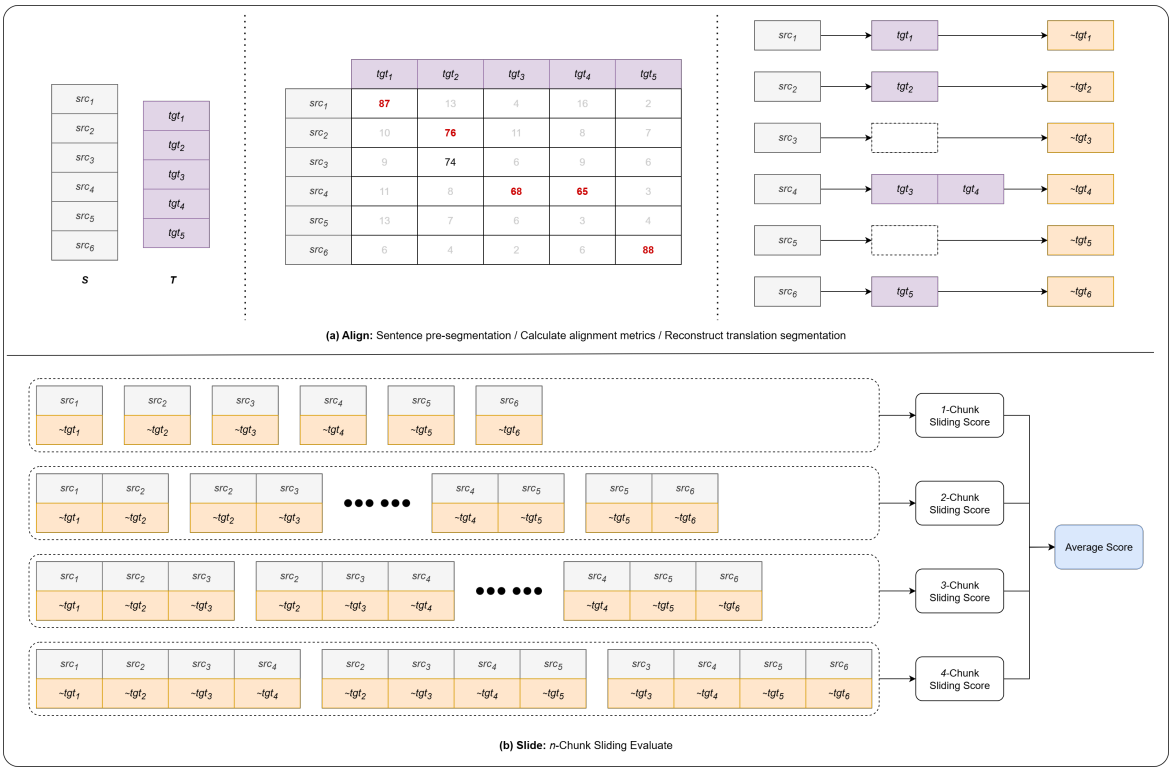

Figure 2: Overall pipeline of Align-then-Slide. (a) Align: constructing a one-to-one sentence-level correspondence between source and translation via optimal dp search. (b) Slide: conducting n-Chunk sliding evaluation with stride one to assess quality at multiple granularities.

challenges:

1. Whole sentence omission breaks the alignment. For example, src_5 is absent from the target.

2. The mapping is no longer bijective: many-to-one (src_2 + src_3 $\rightarrow$ tgt_2) and one-to-many (src_4 $\rightarrow$ tgt_3 + tgt_4) relationships abound;

3. Different systems generate varying numbers of target sentences, shattering the alignment assumption.

In this paper we present ***Align-then-Slide*** (ASD), a complete evaluation framework for ultra-long document-level machine translation. The approach unfolds in two stages.

Stage 1 **Align**: we first pre-segment both source and target texts and compute their sentence-level alignment matrix. Anchoring on the source sentence sequence, we then rebuild the target sequence so that the two have exactly the same number of sentences. During rebuilding, whole-sentence omissions are patched with placeholder sentences and one-to-many mappings are collapsed or expanded, all in one pass. This anchor-on-source design also neutralizes length discrepancies across different systems.

Stage 2 n**-Chunk Sliding Evaluate**: extending the spirit of BLEU's n-grams to SLIDE(Raunak et al., 2024), we compute metrics for 1-, 2-, 3-, and 4-chunk spans and average them. The 1-chunk scores sharply expose omissions, while 2- to 4-chunk scores mitigate the impact of many-to-one mappings, yielding a comprehensive, multi-granularity assessment.

We evaluate Align-then-Slide on the official WMT test suite, where its system-level ranking achieves a Pearson correlation of 0.929 with expert-based MQM(Freitag et al., 2021) scores, confirming its validity. Because this benchmark is already sentence-aligned and free of omissions, we further construct a realistic testbed by using various size LLMs to translate full documents and having professional translators rank the outputs. Align-then-Slide again shows strong agreement with human rankings. Moreover, the preference pairs generated by our metric can be fed directly into CPO (Xu et al., 2024) training, or the metric itself can serve as a reward model for GRPO(Shao et al., 2024). Human evaluation reveals that both CPO- and GRPO-trained systems outperform a vanilla

Algorithm 1 Stage 1: Align

Require: source document S, target document T, source language src_lang, target language tgt_lang
Ensure: aligned source sentences src_lines, reconstructed target sentences new_tgt_lines

// 1. Sentence pre-segmentation
$src\_lines \leftarrow SEG(S, src\_lang)$
$tgt\_lines \leftarrow SEG(T, tgt\_lang)$
$m \leftarrow |src\_lines|$; $n \leftarrow |tgt\_lines|$
// 2. Build $m \times n$ similarity matrix
for $i = 0 \ldots m-1$ **do**
  for $j = 0 \ldots n-1$ **do**
    $score[i][j] \leftarrow EVAL(src\_lines[i], tgt\_lines[j])$ {e.g., COMET-*Kiwi* or LaBSE}
  end for
end for
// 3. Find optimal alignment path via DP
$path \leftarrow DP\_SEARCH(score)$ {Returns list of (i, j) pairs}
// 4. Reconstruct target aligned to source
$new\_tgt\_lines \leftarrow [\,]$
for $i = 0 \ldots m-1$ **do**
  $matched \leftarrow \{j \mid (i, j) \in path\}$
  if $matched = \emptyset$ **then**
    $new\_tgt\_lines[i] \leftarrow$ "" {Placeholder for omission}
  else
    $new\_tgt\_lines[i] \leftarrow CONCAT(tgt\_lines[j]$ **for** j **in** $matched)$
  end if
end for

return src_lines, new_tgt_lines

SFT (supervised fine-tuning) baseline, underscoring the reliability of our evaluation framework.

2 Align-then-Slide

In this paper, we present ***Align-then-Slide*** (ASD), a comprehensive framework for evaluating ultra-long document-level machine translation. The method proceeds in two stages: Stage 1, ***Align***, establishes sentence-level correspondence between source and translation; Stage 2, ***n-Chunked Sliding Evaluate***, performs quality evaluation at multiple granularities.

2.1 Align

In this stage, we automatically segment the entire source and translated documents into a one-to-one set of aligned sentence pairs, as defined in Algorithm 1. The procedure is as follows:

- Sentence pre-segmentation: independently segment both original and translated documents into sentence sequences.
- Calculate alignment metrics: compute sentence-level alignment similarity using reference-free metrics such as COMET-*Kiwi*(Rei et al., 2020b) or LaBSE(Feng et al., 2022).
- Reconstruct translation segmentation: anchored on the source sentence order, we apply a dynamic-programming algorithm to merge or insert placeholder sentences, yielding a target sequence that exactly matches the source in length.

As shown in Figure 2(a), for a source document S and its translation T, we first segment both into sentences using off-the-shelf tools such as *spaCy* and *ersatz*, yielding m source sentences $S = \{src_1, src_2, ..., src_m\}$ and n target sentences $T = \{tgt_1, tgt_2, ..., tgt_n\}$. We then construct an mn similarity matrix, populated by reference-free metrics like COMET-*Kiwi* or LaBSE. When $m = n$ and the mapping is one-to-one, the diagonal attains the maximum scores; in practice $m \neq n$, yet we can still find an optimal path that assigns each source sentence its best-matched target sentence. From the source perspective, unmatched positions are padded with placeholder sentences, while multiple matches are merged, resulting in a reconstructed target sequence $\{\tilde{tgt}_1, \tilde{tgt}_2, \ldots, \tilde{tgt}_m\}$ of identical length m. Finding this optimal path is formulated as a dynamic programming problem as following:

Abstract problem description An $[m, n]$ matrix with values at each point. Starting from $(0, 0)$ and ending at $(m-1, n-1)$, the path requirements are: the y-position must increase by *1* each move, and the x-position must increase by a non-negative number each move, resulting in n points. The goal

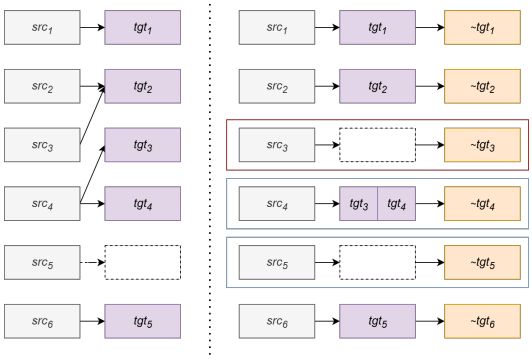

Figure 3: Alignment results after Stage 1. Sentence omissions (e.g., src_5) and one-to-many mappings (e.g., src_4) are resolved, whereas many-to-one conflicts (e.g., src_3 vs. tgt_2) remain and are addressed in Stage 2.

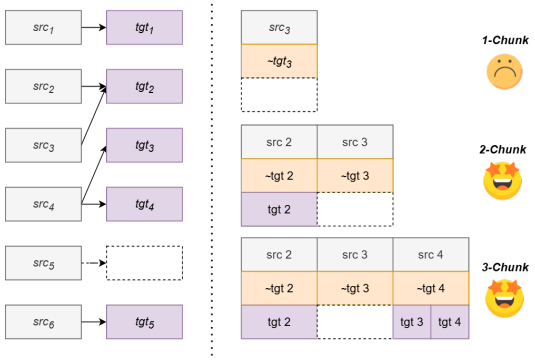

Figure 4: Multi-granularity n-chunk evaluation. While the 1-chunk unit flags src_3 as missing, 2-, 3-, and 4-chunk windows merge adjacent source sentences and restore correct alignment with tgt_2.

is to find a path that maximizes the sum of these n points' values and provides their coordinates. We can use dynamic-programming (*dp*) algorithm to achieve the goal. The algorithm steps are:

- Initialize a 2-D array dp, where $dp[i][j]$ represents the maximum path value sum from $(0, 0)$ to $(m-1, n-1)$.
- Traverse the matrix. For each point (i, j), calculate $dp[i][j]$. Given the y-position must increase by 1 and the x-position by a non-negative number each move, $dp[i][j]$ is the maximum value from $dp[i-1][j-1], dp[i-1][j-2], ..., dp[i-1][0]$, plus the current point's value.
- Finally, $dp[m-1][n-1]$ is the maximum path value sum. Backtracking dp gives the path's point coordinates.

As shown in Figure 2(a), the optimal alignment path via *dp* algorithm is highlighted in red: [(src_1, tgt_1), (src_2, tgt_2), (src_4, tgt_3), (src_4, tgt_4), (src_6, tgt_5)]. Anchoring on the source sequence, we insert empty strings for missed sentences and concatenate multiple targets where necessary, resulting in the final mapping: [(src_1, tgt_1), (src_2, tgt_2), (src_3, ""), (src_4, tgt_3+tgt_4), (src_5, ""), (src_6, tgt_5)]. The reconstructed target sequence $\{\tilde{tgt}_1, \tilde{tgt}_2, \tilde{tgt}_3, \tilde{tgt}_4, \tilde{tgt}_5, \tilde{tgt}_6\}$ is therefore $\{tgt_1, tgt_2, "", tgt_3 + tgt_4, "", tgt_5\}$.

As shown in Figure 3, Stage 1 *Align* successfully resolves omissions (e.g., src_5) and one-to-many mappings (e.g., src_4). Yet many-to-one mappings introduce conflicts: src_3 should share tgt_2 but is instead marked as missing. Merging source sentences seems natural, yet it yields inconsistent segmentations across systems and undermines fair evaluation. We therefore defer this issue to Stage 2, where multi-granularity sliding windows neutralize the conflict.

2.2 n-Chunk Sliding Evaluate

Once sentence-level alignment is established, we conduct a multi-granularity sliding-window evaluation in this stage as shown in Figure 1(b).

Consecutive k sentences form a chunk, and the window slides with a fixed stride of 1, to heighten sensitivity to omissions. As shown in Figure 2(b), for m source sentences $S = \{src_1, src_2, ..., src_m\}$ and their aligned translations $T = \{\tilde{tgt}_1, \tilde{tgt}_2, ..., \tilde{tgt}_m\}$, we set $k \in \{1, 2, 3, 4\}$, yielding $m - k + 1$ units per setting; each unit is scored by a quality estimator, and the scores are averaged to produce the final metric score.

Why n-Chunk? Stage 1 *Align* cannot handle many-to-one mappings: src_3 in Figure 3 is falsely judged empty because it shares tgt_2, causing 1-chunk scores to plummet. Merging source sentences directly would yield inconsistent segmentations across systems. We therefore shift the "merging" into the evaluation stage: when chunk>1, adjacent source sentences are grouped into a single unit, allowing the translation to be re-matched at coarser granularities. As shown in Figure 4, src_3 is correctly aligned within 2-, 3-, and 4-chunk units.

The key distinction from SLIDE lies in our introduction of a hierarchical chunk-based evaluation strategy coupled with a fixed sliding stride of 1, whereas SLIDE employs a dynamic sliding window.

3 Experiments

We conduct two sets of experiments to validate *Align-then-Slide*.

- **Correlation study:** we measure the agreement between our metric and human judgments on both the standard WMT test set and our newly curated real-world benchmark.
- **Training efficacy study:** we fairly compare the quality of translations produced by vanilla supervised fine-tuning (SFT) versus those refined via reinforcement learning (e.g. GRPO) guided by our evaluation strategy.

Our basic settings are as follows. For Stage 1, sentence pre-segmentation is performed with *spaCy*[1], and alignment scores are computed using COMET*Kiwi*. For Stage 2, we instantiate three variants by plugging different COMET backbones, namely COMET20[2], COMET22[3], and COMET*Kiwi*[4], denoted ASD20, ASD22, and ASD*Kiwi*, respectively.

We also experiment with *ersatz*(Wicks and Post, 2021) for pre-segmentation and LaBSE for alignment; these ablations are reported in Section 5.

3.1 Correlation Study Setup

Standard Testsets We adopt the WMT 2020 Chinese→English (ZH→EN) track as our standard benchmarks. This track contains translation submissions for 2000 documents from multiple participating teams. We re-assemble the sentence-level outputs into full documents and evaluate them with *Align-then-Slide*, yielding a system ranking. The resulting system rankings are compared with expert-based MQM rankings via Pearson correlation.

Real-world Testsets Standard test sets are sentence-aligned and therefore do not reflect authentic document-level translations. We therefore constructed new Chinese→English (ZH→EN) and English→Chinese (EN→ZH) test sets, each containing outputs from six Qwen[5] LLMs; construction details are provided in Appendix A. Professional translators produced pairwise relative rankings of these model outputs, establishing human system ranks. We then applied Align-then-Slide to the same outputs to obtain automatic ranks and report the Pearson correlation between the two.

Multilingual Testsets The WMT 2025 Segment-Level Quality Score Prediction shared task[6] provides 16 language-pair test suites whose human ESA or MQM scores serve as the gold standard. Covering typologically diverse directions, including Czech→German (CS→DE), Czech→Ukrainian (CS→UK), English→Arabic (EN→AR), English→Bhojpuri (EN→BHO), English→Chinese (EN→ZH), English→Czech (EN→CS), English→Estonian (EN→ET), English→Icelandic (EN→IS), English→Italian (EN→IT), English→Japanese (EN→JA), English→Maasai (EN→MAS), English→Russian (EN→RU), English→Serbian (EN→SR), English→Ukrainian (EN→UK), English→Korean (EN→JA) and Japanese→Chinese (JA→ZH). These suites offer a large-scale, multilingual benchmark for evaluating quality-estimation metrics. We therefore adopt them as our Multilingual Testsets to examine the cross-lingual stability of Align-then-Slide.

3.2 Training Efficacy Study Setup

Post-training of LLMs typically follows two paradigms: supervised fine-tuning (SFT) and reinforcement learning (RL). For *doc*-mt, the literature has almost exclusively adopted SFT, as parallel document pairs can be readily distilled from state-of-the-art models. While a handful of studies have explored RL, they remain confined to sentence-level tasks, where quality estimation is mature. Owing to the absence of reliable document-level metrics, RL training for full-document translation has been largely unexplored.

We benchmark SFT against RL on the Qwen2.5-7B backbone for both Chinese→English (ZH→EN) and English→Chinese (EN→ZH)to verify the training utility of *Align-then-Slide*.

Data We collect 50k document-level bilingual pairs D. Owing to SFT's sensitivity to quality, we first distill the corpus with Qwen3-32B and Qwen2.5-72B. We then randomly pick one distilled translation per source to create D_{shuf}, and select the higher-scoring translation via *Align-then-Slide* to obtain D_{best}.

SFT Training We directly train two SFT models $M_{sft-shuf}$ and $M_{sft-best}$ on D_{shuf} and D_{best}.

[1] https://spacy.io/models
[2] https://huggingface.co/Unbabel/wmt20-comet-da
[3] https://huggingface.co/Unbabel/wmt22-comet-da
[4] https://huggingface.co/Unbabel/wmt22-cometkiwi-da
[5] https://huggingface.co/Qwen
[6] https://www2.statmt.org/wmt25/mteval-subtask1.html

System	MQM	COMET20	d-COMET20	LLM-as-judge	ASD20
VolcTrans	5.03(1)	0.509(3)	0.366(1)	0.754(2)	0.490(2)
Wechat_AI	5.13(2)	0.522(1)	0.365(2)	0.773(1)	0.496(1)
Tencent	5.19(3)	0.511(2)	0.353(4)	0.754(3)	0.487(3)
OPPO	5.20(4)	0.500(5)	0.331(7)	0.752(5)	0.482(5)
THUMT	5.34(5)	0.497(6)	0.363(3)	0.751(6)	0.485(4)
DeepMind	5.41(6)	0.493(7)	0.346(5)	0.753(4)	0.480(6)
DiDiNLP	5.48(7)	0.502(4)	0.345(6)	0.737(7)	0.477(7)

Table 1: Detailed scores and rankings for seven systems on the WMT2020 Chinese→English test set

RL Training We perform reinforcement learning with Group Relative Policy Optimization (GRPO) (Shao et al., 2024), using *Align-then-Slide* as the reward model and sampling eight translations per step to obtain the final model M_{grpo}.

Evaluation All models are evaluated on the real-world test set. Three professional translators conducted independent, blind scoring; the average ratings produced the human ranking.

We expect two outcomes: SFT on data selected by Align-then-Slide yields higher quality than vanilla SFT baselines; RL training guided by Align-then-Slide surpasses the best SFT result.

4 Results and Analysis

4.1 Correlation Study Results

4.1.1 Results for Standard Testsets

	Pearson	Kendall
MQM	1	1
COMET20*	0.679	0.524
d-COMET20	0.714	0.619
LLM-as-judge	0.857	0.714
ASD20	0.929	0.810

Table 2: Correlation between different rankings and official MQM rankings on the WMT2020 Chinese→English test set for seven systems, measured by Pearson and Kendall.

For the translation outputs of seven systems on the WMT2020 Chinese→English test set, Table 2 presents the correlation of various ranking methods with the MQM rankings. Our *Align-then-Slide* method is benchmarked against other methods such as COMET20, d-COMET20, and an LLM-as-judge (Kocmi and Federmann, 2023) approach, which utilizes the DeepSeek-R1(DeepSeek-AI, 2025) model for document-level scoring. The correlation of various ranking methods with the MQM rankings are quantified by both *Pearson* and *Kendall*. Table 1 presents the detailed scores. All MQM scores and ranks for the seven systems are taken from the official WMT release[7].

Table 2 shows that ASD20 correlates strongly with MQM, achieving a Pearson coefficient of 0.929 and a Kendall's of 0.81, respectively. These scores significantly surpass those of all other systems. Notably, the sentence-level COMET20 method, which operates without document context, demonstrates the weakest correlation — a result to be expected. The document-level COMET20 method, constrained by its use of only a three-window context, shows improvement over the sentence-level version but is still outperformed by ASD20. Owing to its advanced comprehension and reasoning capabilities, the DeepSeek-R1 model delivers performance that is only slightly lower than that of ASD20.

However, the potential variability in LLM-based scoring remains a known challenge. Although we mitigated this by using the mean score from multiple evaluations as the final judgment, the rankings generated by the LLM-as-judge system may still exhibit some randomness. Consequently, we have included the full set of results from the LLM-as-judge system, as well as the precise prompt used for scoring, in the Appendix B and Appendix C for reference.

Table 1 reveals that the two rankings differ only at minor positions, VolcTrans and Wechat_AI swap the 1st and 2nd spots, while OPPO and THUMT exchange the 4th and 5th.

We substituted the evaluation with ASD22 and ASDKiwi and repeated the experiments, again obtaining similar and consistent outcomes, see Appendix D. These findings demonstrate the broad applicability of the *Align-then-Slide* framework.

[7]https://github.com/google/wmt-mqm-human-evaluation

4.1.2 Results for Real-world Testsets

System	Rank	ASD20
Qwen3-32B	1	0.5203(1)
Qwen2.5-72B	2	0.5181(2)
Qwen3-8B	3	0.5041(4)
Qwen2.5-32B	4	0.5096(3)
Qwen2.5-14B	5	0.4939(5)
Qwen2.5-7B	6	0.4906(6)

Table 3: Agreement between ASD20 rankings and human rankings for six LLMs on the Real-world Chinese→English Testsets.

Table 3 compares the rankings produced by *Align-then-Slide* with human judgments across six LLMs on our Real-World Chinese→English Testsets. Remarkably, ASD20 aligns almost perfectly with the human order, misplacing only two systems, and attains a Pearson correlation of 0.943. This provides strong additional evidence for the validity of *Align-then-Slide* in *doc*-mt evaluation.

We repeated the experiment on the Real-world Testsets with ASD22 and ASD*Kiwi*, and further extended it to English→Chinese. All runs achieved similarly high agreement with human rankings. Details are in Appendix D and Appendix E. These results underscore the universality of *Align-then-Slide*.

4.1.3 Results for Multilingual Testsets

Table 4 reports Pearson coefficients for all 16 language directions in the WMT2025 segment-level quality-score prediction track (higher is better).

Our Align-then-Slide ranks first everywhere: except for EN→MAS where it ties the best competitor at 0.597, it outperforms the second-best system in the remaining 15 directions, with the largest margin of 0.502 on EN→RU. The overall average Pearson is 0.775, surpassing the best rival mean by 0.192, further confirming the cross-lingual generality and robustness of our approach.

Directions	Ours	Best of Others
CS→DE	**0.742**	0.65
CS→UK	**0.782**	0.635
EN→AR	**0.855**	0.754
EN→BHO	**0.932**	0.829
EN→CS	**0.696**	0.609
EN→ET	**0.745**	0.686
EN→IS	**0.793**	0.74
EN→IT	**0.717**	0.422
EN→JA	**0.767**	0.467
EN→KO	**0.773**	0.56
EN→MAS	**0.597**	**0.597**
EN→RU	**0.835**	0.333
EN→SR	**0.903**	0.829
EN→UK	**0.819**	0.419
EN→ZH	**0.747**	0.473
JA→ZH	**0.698**	0.318
AVG	0.775	0.583

Table 4: Results in all 16 language directions of the Segment-level quality score prediction track at WMT2025.

4.2 Training Efficacy Study Results

Accordingly, we set two validation goals:

- At the data level, to demonstrate that Align-then-Slide can reliably pick "the best of the best"—SFT on D_{best}, which contains only the highest-scoring distilled translations per source, should significantly outperform vanilla SFT baselines trained on single-model distillations or shuffle data D_{shuf};
- At the training-paradigm level, to provide the first practical evidence that document-level RL is viable—using Align-then-Slide as the reward signal and GRPO for online policy optimisation, the resulting model should surpass the strongest SFT baseline, breaking the current "doc-mt can only be fine-tuned" mindset and establishing RL as a usable post-training strategy for full-document translation.

Model	ZH→EN	EN→ZH
Baseline	83.1(4)	82.7(4)
$M_{sft-shuf}$	86.4(3)	84.9(3)
$M_{sft-best}$	88.7(2)	86.3(2)
M_{grpo}	90.2(1)	87.2(1)

Table 5: Human evaluation rankings on the Chinese→English and English→Chinese test sets.

This demonstrates two key points. First, RL training guided by *Align-then-Slide* significantly outperforms all SFT baselines, confirming the framework's effectiveness for document-level RL. Second, among the SFT models, the one trained on D_{best}, selected via Align-then-Slide, achieves the highest quality, showing that D_{best} is consistently superior to $D_{sft-shuf}$. Thus, Align-then-Slide not only steers training but also reliably identifies higher-quality data, underscoring its utility

System	Expert Rank	spaCy + COMET*Kiwi*	ersatz + COMET*Kiwi*	spaCy + Labse
Qwen3-32B	1	0.5203(1)	0.5201(1)	0.5203(1)
Qwen2.5-72B	2	0.5181(2)	0.5180(2)	0.5181(2)
Qwen3-8B	3	0.5041(4)	0.5041(4)	0.5041(4)
Qwen2.5-32B	4	0.5096(3)	0.5095(3)	0.5096(3)
Qwen2.5-14B	5	0.4939(5)	0.4936(6)	0.4940(5)
Qwen2.5-7B	6	0.4906(6)	0.4911(5)	0.4906(6)

Table 6: Ablation results on segmentation tools (*spaCy* vs. *ersatz*) and similarity metrics (COMET*Kiwi* vs. LaBSE) within the alignment stage of Align-then-Slide.

across the entire training pipeline.

5 Ablation Study

In this section, we examine how different sentence-segmentation tools and different alignment models affect the final evaluation results.

5.1 Setup

For Stage 1, sentence pre-segmentation is performed with *spaCy*[8], and alignment scores are computed using COMET*Kiwi*. In Stage 2, we instantiate three variants by plugging different COMET backbones, namely COMET20[9], COMET22[10], and COMET*Kiwi*[11], denoted ASD20, ASD22, and ASD*Kiwi*, respectively. We also experiment with *ersatz*(Wicks and Post, 2021) for pre-segmentation and LaBSE for alignment; these ablations are reported in Section 5.

5.2 Analysis

Table 6 ablates the two pivotal components of the alignment stage: the document pre-segmentation tool and the model that produces the m×n similarity matrix. We compare *spaCy* and *ersatz* for segmentation, and COMET*Kiwi* and LaBSE for alignment modeling, all introduced in Section 5.1.

The three columns of Table 6 correspond to these configurations: the original "*spaCy* + COMET*Kiwi*", the segmentation variant "*ersatz* + COMET*Kiwi*", and the similarity-metric variant "*spaCy* + LaBSE".

Segment Tools Comparing the "*spaCy* + COMET*Kiwi*" and "*ersatz* + COMET*Kiwi*" columns reveals that ASD20 scores for each system differ only marginally, and the predicted rankings remain identical. This stability stems from the maturity of current segmentation tools: a spot-check of 120 sentences shows just seven (5.83%) segmentation mismatches. Because both hypotheses and references are aligned to the source segmentation, these minor tool differences have negligible impact on the final evaluation.

Alignment Models After segmentation, alignment hinges solely on the $m \times n$ similarity matrix. Comparing "*spaCy* + COMET*Kiwi*" and "*spaCy* + LaBSE" shows near-identical ASD20 scores and identical rankings across systems. This stems from both the high accuracy of current alignment models and our DP algorithm's ability to find a consistently optimal path, underscoring the robustness of the ASD approach.

As shown, our framework is highly robust: varying the segmentation tool or alignment model produces negligible differences in the final evaluation scores.

6 Conclusion

Targeting *doc*-mt evaluation challenges, this paper proposes an align-then-slide evaluation method, forming a complete metric system. By automatically constructing sentence-level alignment and combining it with n-chunk sliding evaluation, it overcomes traditional metric limitations and offers a full solution for *doc*-mt evaluation. Future work will focus on further algorithm optimization to enhance the metrics' accuracy and efficiency.

7 Limitations

Computational Cost. Generating the $m \times n$ similarity matrix and performing sliding-window evaluation incurs $\mathcal{O}(m \times n)$ memory and $\mathcal{O}(k \times (m - k + 1))$ extra calls to the backbone model. For very long documents, GPU memory and latency can become prohibitive without batching or pruning heuristics.

[8] https://spacy.io/models
[9] https://huggingface.co/Unbabel/wmt20-comet-da
[10] https://huggingface.co/Unbabel/wmt22-comet-da
[11] https://huggingface.co/Unbabel/wmt22-cometkiwi-da

References

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. *Preprint*, arXiv:2007.01852.

Patrick Fernandes, Kayo Yin, Graham Neubig, and André F. T. Martins. 2021. Measuring and increasing context usage in context-aware machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6467–6478. 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.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

Jiaxin GUO, Xiaoyu Chen, Zhiqiang Rao, Jinlong Yang, Zongyao Li, Hengchao Shang, Daimeng Wei, and Hao Yang. 2025. Automatic evaluation metrics for document-level translation: Overview, challenges and trends. *Preprint*, arXiv:2504.14804.

Yunsu Kim, Duc Thanh Tran, and Hermann Ney. 2019. When and why is document-level context useful in neural machine translation? In *Proceedings of the Fourth Workshop on Discourse in Machine Translation, DiscoMT@EMNLP 2019, Hong Kong, China, November 3, 2019*, pages 24–34. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Gemba-mqm: Detecting translation quality error spans with gpt-4. In *Proceedings of the Eighth Conference on Machine Translation*, Singapore. Association for Computational Linguistics.

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2022. A survey on document-level neural machine translation: Methods and evaluation. *ACM Comput. Surv.*, 54(2):45:1–45:36.

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, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

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: Short Papers, NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 205–211. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020a. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 2685–2702. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020b. Unbabel's participation in the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation, WMT@EMNLP 2020, Online, November 19-20, 2020*, pages 911–920. Association for Computational Linguistics.

Ricardo Rei, Marcos V. Treviso, Nuno Miguel Guerreiro, Chrysoula Zerva, Ana C. Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte M. Alves, Luísa 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 2022, Abu Dhabi, United Arab Emirates (Hybrid), December 7-8, 2022*, pages 634–645. Association for Computational Linguistics.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

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 2022, Abu Dhabi, United Arab Emirates (Hybrid), December 7-8, 2022*, pages 118–128. 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.

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. *Preprint*, arXiv:2401.08417.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.

A Appendix: Real-world Testsets

Our bilingual data originate from CommonCrawl. We first randomly sampled 100 document pairs that contained both source and target texts. After rule-based filtering to remove poorly aligned samples, professional translators selected the 50 highest-quality pairs to form our real-world test set.

System	MQM	COMET20	d-COMET20	LLM-as-judge_1	LLM-as-judge_2	ASD20
VolcTrans	5.03(1)	0.509(3)	0.366(1)	0.754(2)	0.756(2)	0.490(2)
Wechat_AI	5.13(2)	0.522(1)	0.365(2)	0.773(1)	0.764(1)	0.496(1)
Tencent	5.19(3)	0.511(2)	0.353(4)	0.754(3)	0.748(5)	0.487(3)
OPPO	5.20(4)	0.500(5)	0.331(7)	0.752(5)	0.749(4)	0.482(5)
THUMT	5.34(5)	0.497(6)	0.363(3)	0.751(6)	0.752(3)	0.485(4)
DeepMind	5.41(6)	0.493(7)	0.346(5)	0.753(4)	0.745(6)	0.480(6)
DiDiNLP	5.48(7)	0.502(4)	0.345(6)	0.737(7)	0.741(7)	0.477(7)

Table 7: Detailed scores and rankings for seven systems on the WMT2020 Chinese→English test set

B Appendix: Full set of results

	Pearson	Kendall
MQM	1	1
COMET20*	0.679	0.524
d-COMET20	0.714	0.619
LLM-as-judge_1	0.857	0.714
LLM-as-judge_2	0.821	0.619
ASD20	0.929	0.810

Table 8: Correlation between different rankings and official MQM rankings on the WMT2020 Chinese→English test set for seven systems, measured by Pearson and Kendall.

C Appendix: The precise prompt used for scoring

You are a professional translation quality assessment expert. Please rate the following system translation.

Assessment Task:

Rate the translation quality of the system output on a scale of 0 to 100 based on the original text and reference translation.

Input Information:

1. Original Text: [src]
2. Reference Translation: [tgt]
3. System Translation: [mt]

Scoring Criteria (Out of 100):

- Accuracy (40 points):
Whether the translation faithfully conveys the meaning of the original text.
Reference dimensions: mistranslation, over-translation, omission, logic, etc.

- Fluency (25 points):
Whether the translation is natural and fluent in the target language and conforms to linguistic conventions.
Reference dimension: fluency, etc.

- Style Matching (10 points):
Whether the translation matches the style and tone of the reference translation.

- Terminology Consistency (25 points):
Whether the translation of professional terms and key concepts is accurate and consistent.
Reference dimensions: NE, terminology, etc.

Output Requirements:

Please strictly follow the format below for your output, without any additional content:
Translation Quality Score: an integer score between 0 and 100

Please begin the assessment.

Table 9: The precise prompt used for scoring

D Appendix: Results for Standard Testsets

	Pearson	Kendall
ASD20	0.929	0.810
ASD22	0.893	0.714
ASD*Kiwi*	0.964	0.905

Table 10: Correlation between ASD20 / ASD22/ ASD*Kiwi* rankings and official MQM rankings on the WMT2020 test set for seven systems, measured by Pearson and Kendall.

E Appendix: Results for Real-world Testsets

System	**Rank**	**ASD20**	**ASD22**	**ASD*Kiwi***
Qwen3-32B	1	0.5203(1)	0.8378(2)	0.8204(2)
Qwen2.5-72B	2	0.5181(2)	0.8385(1)	0.8207(1)
Qwen3-8B	3	0.5041(4)	0.8362(3)	0.8190(4)
Qwen2.5-32B	4	0.5096(3)	0.8345(4)	0.8192(3)
Qwen2.5-14B	5	0.4939(5)	0.8304(5)	0.8187(5)
Qwen2.5-7B	6	0.4906(6)	0.8293(6)	0.8184(6)

Table 11: Agreement between ASD20 / ASD22/ ASD*Kiwi* rankings and human rankings for six LLMs on the Real-world Chinese→English Testsets.

System	**Rank**	**ASD20**	**ASD22**	**ASD*Kiwi***
Qwen3-32B	1	0.3712(2)	0.7261(1)	0.8245(2)
Qwen2.5-72B	2	0.3746(1)	0.7255(2)	0.8287(1)
Qwen2.5-32B	3	0.3502(3)	0.7125(4)	0.8189(4)
Qwen3-8B	4	0.3486(4)	0.7134(3)	0.8203(3)
Qwen2.5-14B	5	0.3361(5)	0.7018(5)	0.8138(5)
Qwen2.5-7B	6	0.3144(6)	0.6910(6)	0.7325(6)

Table 12: Agreement between ASD20 / ASD22/ ASD*Kiwi* rankings and human rankings for six LLMs on the Real-world English→Chinese Testsets.