STAR : Sentence Translation Alignment Rate for Document-to-Document Machine Translation

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have enabled a shift from sentence-level to document-to-document (Doc2Doc) machine translation, promising improved global coherence. However, document-to-document generation in a single pass frequently suffers from structural misalignment, manifesting as sentence omissions or hallucinations that violate the core requirement of source-target correspondence. To address this, we introduce **S**entence **T**ranslation **A**lignment **R**ate (**STAR**), an auxiliary metric that explicitly quantifies sentence-level structural fidelity. Building on this, we propose **STAR**-masked **P**reference **O**ptimization (**StarPO**), a framework that ranks document-level hypotheses by structural quality and utilizes a dynamic alignment mask to focus optimization on misaligned segments. Experimental results across news and literary domains demonstrate that StarPO significantly enhances translation quality and structural integrity. Notably, StarPO allows compact models (e.g., Qwen3-4B) to surpass the performance of massive proprietary systems like GPT-4o while maintaining superior token efficiency. We will release our code and datasets.

1 Introduction

Recent advances in large language models (LLMs) has significantly advanced document-level machine translation (DocMT) (Wang et al., 2023, 2025f; Karpinska and Iyyer, 2023). With long-context modeling and strong generative capabilities, LLMs make it increasingly feasible to move beyond sentence-to-sentence (Sent2Sent) translation toward document-to-document (Doc2Doc) translation, where an entire source document is translated in a single pass. Despite this promise, Doc2Doc translation often exhibits sentence-level structural failures, such as sentence omissions or hallucinations, that violate the core requirement of sentence-level correspondence between source and target documents (Dong et al., 2025). In this work, we aim to improve Doc2Doc translation by addressing sentence-level structural misalignments, so that better alignment directly contributes to higher translation quality.

System	Ideal	Structural Deviations		
	1-to-1	1-to-0	0-to-1	Other
LLaMA-3.1-8B	92.59	2.08	0.17	4.15
Qwen-2.5-7B	95.35	1.91	1.31	1.43
Qwen3-4B	94.72	0.72	0.33	4.23
Deepseek-R1	95.03	4.85	0.03	0.09
GPT-4o	92.91	2.25	2.89	1.95
Prompting LLaMA-3.1-8B *with Sentence Splitter*				
MixSFT (Li et al., 2026)	96.78	0.57	2.08	0.57
KFMT (Liu et al., 2025)	95.48	1.89	2.34	0.29
Ours (on Qwen3-4B)	98.09	0.64	0.00	1.27

Table 1: Alignment distribution (in %) for in-one-go Doc2Doc Chinese-to-English translation on News-Commentary, computed by Gemini-2.5-Flash.

Most existing DocMT approaches implicitly avoid structural misalignment by adopting sentence-level or context-aware paradigms, where translation is still performed at the sentence level with surrounding context (Lyu et al., 2024; Hu et al., 2025; Wu et al., 2024a). While this design enforces sentence correspondence, it usually limits its global planning and incurs substantial computational overhead, as source sentences are repeatedly encoded across overlapping contexts.[1]

Ideally, Doc2Doc translation preserves a clear correspndence between source and target sentences. In practice, howerver, model outputs often contain 1-to-0 alignments (omissions) or 0-to-1 alignments (hallucinations).[2] As shown in our

[1] See Appendix A for additional discussions.

[2] Human translations do not always preserve strict 1-to-1 sentence correspondence, for example, two source sentences may merge into one target sentence, but this paper focuses on mitigating clear structural misalignments such as omissions or hallucinations rather than enforcing rigid 1-to-1 mapping.

preliminary analysis in Table 1, such errors are pervasive across model scales and persist even when sentence boundary constraints are explicitly introduced during generation (Liu et al., 2025; Li et al., 2026). Notably, even strong proprietary models such as GPT-4o (OpenAI, 2024) exhibit nontrivial rates of structural mismatch in Doc2Doc settings (Liu et al., 2024; Shao et al., 2024a).

These errors are particularly pronounced in long or complex documents and closely related to over-rejection (Xu et al., 2024, 2025) and reward hacking (Shihab et al., 2025), where models adopt overly conservative strategies under safety or short-context biases. Importantly, structural misalignment is largely invisible to standard training objectives and evaluation metrics (e.g. COMET (Rei et al., 2022a)), which primarily focus on semantic adequacy and fluency. As a result, Doc2Doc systems may achieve high metric scores while still omitting or hallucinating content.

To address this limitation, we introduce **S**entence **T**ranslation **A**lignment **R**ate (**STAR**), an auxiliary metric that explicitly measures sentence-level structural fidelity in Doc2Doc translation. Building on STAR, we propose **STAR**-masked **P**reference **O**ptimization (**StarPO**) framework that ranks document-level hypotheses by alignment quality and encourage structurally faithful generation. We further introduce a dynamic alignment masking strategy that downweights or excludes sentences that are already well aligned, allowing optimization to focus on misaligned segments such as omissions and hallucinations.

In summary, our contributions are:

- We identify sentence-level structural misalignment as a key bottleneck in Doc2Doc translation and introduce **STAR** , a novel auxiliary metric for measuring structural fidelity.
- We propose **STAR**-masked **P**reference **O**ptimization (**StarPO**) with dynamic alignment masking to mitigate sentence omissions and hallucinations.
- We demonstrate consistent and robust improvements in document-level translation quality across multiple domains and models.

2 Methodology

In this section, we present a novel framework designed to enhance structural fidelity of document-level translation as illustrated in Figure 1.

2.1 Sentence Translation Alignment Rate (STAR)

We introduce Sentence Translation Alignment Rate (**STAR**), a metric for quantifying sentence-level structural fidelity in Doc2Doc translation. As illustrate in Figure 1(a), the computation of STAR proceeds in four steps:

1. Sentence Segmentation: Source and target documents (S and T) are segmented into sentences, $S = \{s_1, \ldots, s_m\}$ and $T = \{t_1, \ldots, t_n\}$. Specifically, we segment sentences using SaT (Frohmann et al., 2024).

2. Sentence-Level Alignment: Sentences from the source and target are aligned to form disjoint alignment units $\mathcal{U} = \{u_1, \ldots, u_K\}$, where each unit $u_k = (\mathbf{s}_k, \mathbf{t}_k)$ may contain zero or more source sentences $\mathbf{s}_k \subseteq S$ and zero or more target sentences $\mathbf{t}_k \subseteq T$. These units are *minimal* and cannot be further decomposed. Here, we use Bertalign (Liu and Zhu, 2023) for sentence-level alignment.

3. Unit Categorization: Each alignment unit is classified based on the number of source and target sentences: (1) 1-to-1 ($\mathcal{U}_{1:1}$), where $|\mathbf{s}_k| = |\mathbf{t}_k| = 1$; (2) Deletion ($\mathcal{U}_{1:0}$), where $|\mathbf{s}_k| \geq 1, |\mathbf{t}_k| = 0$; (3) Insertion ($\mathcal{U}_{0:1}$), where $|\mathbf{s}_k| = 0, |\mathbf{t}_k| \geq 1$; and (4) Complex ($\mathcal{U}_{\text{complex}}$), covering all other cases ($|\mathbf{s}_k| + |\mathbf{t}_k| > 2$ with $|\mathbf{s}_k|, |\mathbf{t}_k| \geq 1$).

4. STAR Computation: STAR is the fraction of clean 1-to-1 units among all alignment units:

$$\text{STAR}(S, T) = \frac{|\mathcal{U}_{1:1}|}{|\mathcal{U}_{1:1}| + |\mathcal{U}_{1:0}| + |\mathcal{U}_{0:1}| + |\mathcal{U}_{\text{complex}}|}. \quad (1)$$

Thus, higher STAR indicates better sentence-level structural fidelity while lower STAR reflects deletions, insertions, or complex merges. Notably, STAR can also be computed directly via an LLM-as-a-judge approach, based on the four steps outlined above. See Appendix C for detailed prompts.

2.2 Preference Data Generation

To support STAR-masked preference optimization (StarPO), we construct a document-level preference dataset from automatically generated translation candidates. For each source document, we

See Appendix B for detailed case study.

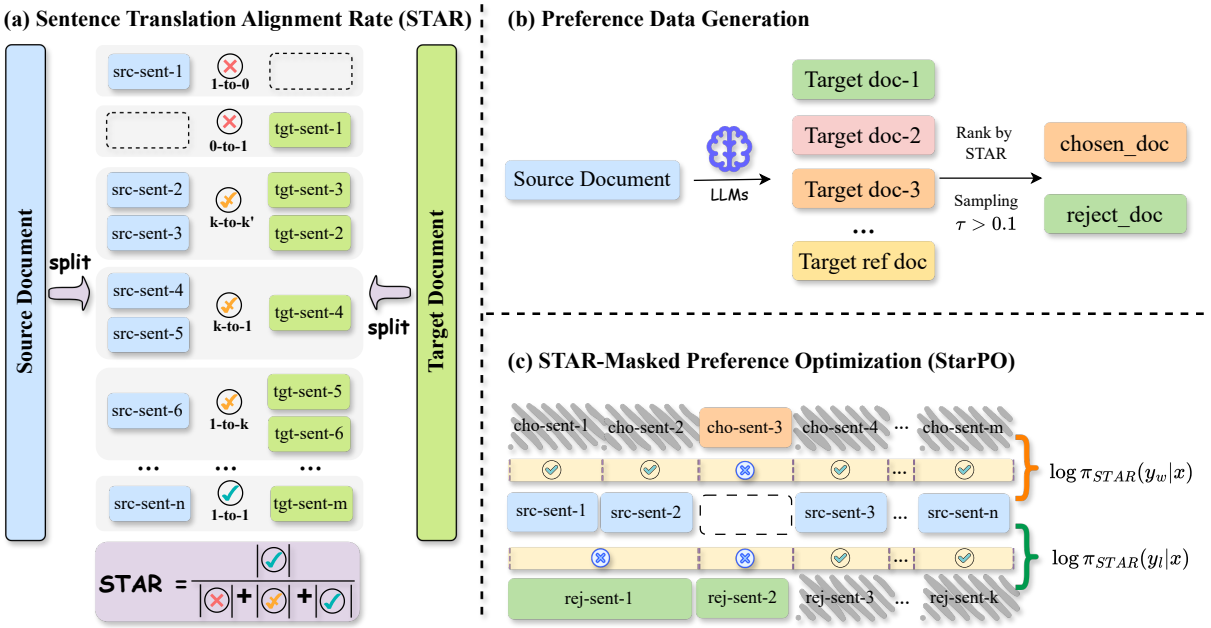

Figure 1: Overview of the proposed framework. (a) Sentence Translation Alignment Rate (STAR): Source and target documents are segmented and aligned to compute STAR, defined as the ratio of strict 1-to-1 alignments. (b) Preference Data Generation: Translation candidates are sampled and ranked by their STAR scores. Pairs exhibiting a score disparity larger than τ are selected as chosen (y_w) and rejected (y_l) samples. (c) STAR-Masked Preference Optimization (StarPO): A sentence-level mask is applied to the CPO objective, excluding well-aligned (1-to-1) sentences to focus optimization exclusively on structurally problematic segments.

use GPT-4o (OpenAI, 2024) to generate 5 diverse translation candidates with a temperature of 1.0. Reference translation without sentence boundaries is also included in the candidate pool when available.

For each souce document S, we compute the STAR score (Section 2.1) for all candidates and form a preference pair by selecting the candidate with the highest STAR score as the *chosen* example (T_w) and the one with the lowest score as the *rejected* example (T_l). To ensure meaningful supervision, a pair (T_w, T_l) is retained only if the STAR score difference exceeds a threshold τ, i.e., $|\text{STAR}(S, T_w) - \text{STAR}(S, T_l)| > \tau$. In our experiments, we set $\tau = 0.1$.

2.3 STAR-Masked Preference Optimization (StarPO)

Following Wang et al. (2025a); Agrawal et al. (2024); Xu et al. (2024); Sun et al. (2025a), we adopt a two-stage training paradigm. We first perform supervised fine-tuning (SFT) on high-quality parallel corpora to establish a warm-started policy π_{SFT}. We then apply preference optimization to further align the model with structural constraints.

Background: Contrastive Preference Optimization. Following contrastive preference optimization (CPO) (Xu et al., 2024, 2025), we assume a preference dataset $\mathcal{D} = \{(S, T_w, T_l)\}$, where S is a source document and (T_w, T_l) denote a preferred and dis-preferred translation, respectively. CPO optimizes the policy π_θ by maximizing the likelihood margin between the preferred and dis-preferred candidates. The standard objective is formulated as:

$$\begin{aligned}\mathcal{L}_{\text{CPO}}(\pi_\theta) = &- \mathbb{E}_{\mathcal{D}}\Big[\log \sigma\big(\beta \log \pi_\theta(T_w|S) \\ &- \beta \log \pi_\theta(T_l|S)\big)\Big] \\ &- \mathbb{E}_{\mathcal{D}}\big[\log \pi_\theta(T_w|S)\big],\end{aligned} \tag{2}$$

where σ is sigmoid function and β is a hyperparameter controlling strength of preference signal. Typically, $\log \pi_\theta(T|S)$ is computed as the sum of log-probabilities over all tokens in document T.

STAR-Masked Objective. Directly applying CPO at the document-level treats all tokens equally, regardless of whether they correspond to structurally correct translations or not. To focus learning on sentence-level structural issues (e.g., hallucinations, omissions, or complex alignments), we introduce a STAR-based masking strategy.

Given a target document T which is segmented into n sentences $T = \{t_1, \ldots, t_n\}$ and the alignment units defined in Section 2.1, we define a sentence-level mask $\mathcal{M}(t_j)$ that indicates whether sentence t_j should contribute to preference loss:

$$\mathcal{M}(t_j) = 1 - \mathbb{I}_{1:1}(t_j), \tag{3}$$

where $\mathbb{I}_{1:1}(t_j) = 1$ if sentence t_j belongs to a clean 1-to-1 alignment unit and 0 otherwise.

As a result, well-aligned sentences receive a mask value of 0, whereas sentences associated with structural mismatches receive a value of 1.

Using this mask, we define STAR-Masked log-likelihood, $\log \pi_{\text{STAR}}(T|S)$, which aggregates token-level probabilities strictly for non-1-to-1 sentences:

$$\log \pi_{\text{STAR}}(T|S) = \sum_{j=1}^{n} \Big[\mathcal{M}(t_j) \cdot \sum_{k=1}^{|t_j|} \log \pi_\theta \left(t_{j,k} | t_{<j}, t_{j,<k}, S \right) \Big], \tag{4}$$

where $|t_j|$ is the number of tokens in sentence t_j, $t_{j,k}$ denotes the k-th token in the t_j, $t_{<j}$ represents all target sentences preceding t_j, and $t_{j,<k}$ denotes the generated tokens in t_j, respectively. The inner summation corresponds to the standard token-level log-likelihood for sentence t_j.

We then replace the standard document-level likelihood in CPO with the masked likelihood to obtain the STAR-masked preference loss:

$$\begin{aligned} \mathcal{L}_{\text{STAR-CPO}}(\pi_\theta) = &- \mathbb{E}_{\mathcal{D}} \Big[\log \sigma \big(\beta \log \pi_{\text{STAR}}(T_w|S) \\ &- \beta \log \pi_{\text{STAR}}(T_l|S) \big) \Big] \\ &- \mathbb{E}_{\mathcal{D}} \big[\log \pi_{\text{STAR}}(T_w|S) \big]. \end{aligned} \tag{5}$$

3 Experimentation

3.1 Experimental Settings

Datasets. Following recent works (Wang et al., 2025f; Liu et al., 2025; Cui et al., 2024), we evaluate on both literary and news domains. Specifically, we use Guofeng dataset (Wang et al., 2024)[3] for Chinese (Zh) ⇔ {English (En), German (De), Russian (Ru)} translation, and News-Commentary v18.1 from WMT25[4] for English (En) ⇔ { Chinese (Zh), German (De), Russion (Ru), Spanish (Es)} and Chinese (Zh) ⇔ German (De). We use the training sets to fine-tune the LLMs and to construct preference pairs. Detailed dataset statistics are provided in Appendix D.

Models and Implementation Details. We use three open-source instruction-tuned large language models: LLaMA-3.1-8B-Instruct (Team, 2024a)[5], Qwen-2.5-7B-Instruct (Team, 2024b)[6], and Qwen-3-4B-Instruct (Yang et al., 2025)[7]. See Appendix E for more implementation details.

Systems. For comprehensive comparison, we evaluate two categories of systems:

(1) Training Paradigms. We report the performance of the base LLMs, models fine-tuned with supervised fine-tuning (**+SFT**), further optimized with standard contrastive preference optimization (**+CPO**), and our proposed STAR-masked preference optimization (**+StarPO**).

(2) State-of-the-Art and Competitive Systems. We additionally compare against strong document-level translation systems, including Tower-plus-9B (Rei et al., 2024), GPT-4o (OpenAI, 2024), and Deepseek-R1 (DeepSeek-AI, 2025).

Metrics. We evaluate translation quality using document-level COMET (dCOMET) (Vernikos et al., 2022), computed by `wmt22-comet-da` (Rei et al., 2022a). Specifically, since document-level evaluation requires sentence-level alignments (Vernikos et al., 2022), we apply the alignment strategy described in Section 2.1. Following Zouhar et al. (2024); Zebaze et al. (2025), we assign a score of 0 to unaligned segments, including omissions (1-to-0) and hallucinations (0-to-1). Importantly, **complex mappings** (1-to-k, k-to-1, k-to-k') **retain** their computed COMET scores. See Appendix F for additional results using COMETKiwi (Rei et al., 2022b) and our proposed STAR scores.

3.2 Main Results

Our main results are shown in Table 2 and Table 3.

Performance on News-Commentary. Table 2 presents the results on the News-Commentary dataset. Across all backbones (LLaMA-3.1, Qwen2.5, and Qwen3), supervised fine-tuning (+SFT) yields consistent and substantial improvement over the base models on most language pairs. Building on this, standard CPO (+CPO) further enhances translation quality. Our proposed StarPO consistently outperforms standard CPO

[3] github.com/longyuewangdcu/GuoFeng-Webnovel/
[4] www2.statmt.org/wmt25/translation-task.html
[5] huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[6] huggingface.co/Qwen/Qwen2.5-7B-Instruct
[7] huggingface.co/Qwen/Qwen3-4B-Instruct-2507

System	Zh ⇔ En		De ⇔ En		De ⇔ Zh		Ru ⇔ En		En ⇔ Es		Avg.
	⇒	⇐	⇒	⇐	⇒	⇐	⇒	⇐	⇒	⇐	
LLAMA-3.1-8B-INSTRUCT											
Base	72.23	74.89	71.60	80.13	72.49	58.80	80.65	80.75	80.53	78.62	75.07
+ SFT	76.81	80.01	83.79	81.47	75.42	75.41	82.37	80.69	80.66	83.49	80.01
+ CPO	81.10	79.94	83.98	82.42	75.50	75.05	82.51	81.43	81.45	84.60	80.80
+ StarPO	81.55	80.11	84.05	82.50	76.33	**77.08**	82.53	81.71	82.03	84.91	81.28
QWEN2.5-7B-INSTRUCT											
Base	81.57	80.40	83.31	79.09	77.60	73.97	81.94	78.35	81.33	84.34	80.19
+ SFT	82.01	80.89	83.61	80.02	77.68	74.45	82.26	81.65	81.75	84.73	80.91
+ CPO	81.94	80.97	84.01	81.14	**78.27**	74.56	82.59	81.79	81.65	84.71	81.16
+ StarPO	**82.27**	**81.33**	**84.06**	81.89	78.22	74.58	**82.70**	82.12	81.79	**85.21**	81.42
QWEN3-4B-INSTRUCT											
Base	81.47	80.31	83.63	80.87	77.43	76.78	81.58	84.03	80.21	84.01	81.03
+ SFT	81.71	80.88	83.65	81.50	77.37	76.87	82.19	84.04	80.46	84.70	81.34
+ CPO	81.77	80.90	83.56	81.55	77.79	76.92	82.23	84.01	81.20	84.67	81.46
+ StarPO	82.24	81.17	83.84	82.07	78.00	76.93	82.43	84.21	81.48	84.93	**81.73**
OTHER SYSTEMS											
Tower+	80.53	80.18	84.01	82.57	75.85	76.34	82.58	81.66	82.46	84.81	81.10
GPT-4o	77.42	80.64	83.87	82.84	77.88	69.32	82.36	84.73	**83.31**	84.83	80.72
Deepseek-R1	79.52	79.10	82.19	**82.90**	77.92	75.96	81.90	**84.97**	82.56	82.71	80.97

Table 2: Performance in dCOMET score on the News-Commentary test set. **Bold** scores represent the global best performance and underlined scores represent the global second-best performance. Blue text background indicates that the improvement over the origin Base model achieves at least 85% accuracy with the human judgment (Xu et al., 2024; Kocmi et al., 2024b). Specifically, the improvement needs a minimin of ≥ 0.71 for `wmt22-comet-da`.

System	Zh ⇔ En		Zh ⇔ De		Zh ⇔ Ru	
	⇒	⇐	⇒	⇐	⇒	⇐
LLAMA-3.1-8B-INSTRUCT						
Base	53.66	64.33	52.31	66.36	60.28	62.04
+ SFT	62.70	65.42	67.96	73.33	75.27	72.75
+ CPO	63.01	68.70	69.04	73.95	75.22	72.92
+ StarPO	63.43	72.17	**72.15**	74.61	**77.50**	73.79
QWEN2.5-7B-INSTRUCT						
Base	70.22	71.99	64.11	73.04	55.74	73.31
+SFT	70.42	69.54	66.66	73.03	75.86	72.23
+CPO	70.20	71.23	66.51	73.69	74.63	72.78
+StarPO	**70.64**	72.13	70.72	73.94	77.46	75.00
QWEN3-4B-INSTRUCT						
Base	63.61	68.64	68.01	73.76	72.98	73.04
+SFT	68.61	69.77	69.41	75.08	76.48	75.06
+CPO	68.77	71.02	69.01	74.84	76.49	73.83
+StarPO	70.13	72.21	70.20	**75.20**	76.53	**75.44**
OTHER SYSTEMS						
Tower+	69.39	65.92	54.41	72.89	70.39	66.48
GPT-4o	69.58	**73.61**	53.24	73.91	55.35	70.65
Deepseek	62.86	69.58	64.51	70.30	71.53	67.35

Table 3: Performance in dCOMET on Guofeng test set.

and achieves the best average performance across models. For instance, on LLaMA-3.1, StarPO obtains an average improvement of 0.48 COMET over standard CPO. Importantly, these gains are stable across model scales. With StarPO, the compact Qwen3-4B model outperforms strong SOTA and competitive systems, including Tower-plus-9B, GPT-4o, and DeepSeek-R1. This highlights the effectiveness of structural alignment preference optimization for document-level translation.

Performance on Guofeng. Results on the Guofeng dataset (Table 3) further highlight the robustness of our approach in a challenging literary domain. It is worth noting that literary translation typically does not adhere to rigid 1-to-1 literalism, often employing complex sentence mappings for stylistic flow. Although StarPO enforces strict 1-to-1 alignment constraints during the training phase, our evaluation protocol (dCOMET) remains inclusive of valid complex mappings. The significant performance gains suggest that this strict training signal does not negatively impact literary flexibility; instead, it effectively mitigates severe generative pathologies - such as language mismatches, omissions, and hallucinations - thereby ensuring the structural integrity required for coherent document translation. For example, the base model of LLaMA-3.1-Instruct fails to generate valid outputs in certain language directions (e.g., 52.31 COMET on Zh⇒De), and even GPT-4o shows limited adaptation to the target style. In contrast, StarPO consistently improves performance across all backbones, yielding substantial gains over both SFT and CPO. With StarPO, Qwen3-4B achieves the best performance in several directions, significantly outperforming GPT-4o (e.g., +15.0 COMET on Zh ⇒ En). These results indicate that smaller models, when trained with structural-alignment preference optimization, can outperform massive generalist models on complex, stylized document-level translation tasks.

Metric	Spearman (ρ)	Pearson (r)	Kendall (τ)
Ground Truth	1.0000	1.0000	1.0000
LLM-based Evaluation			
LLM-judge (Appendix C)	**0.7951**	**0.8321**	**0.8015**
Existing Alignment Methods			
Align-then-Slide	0.4169	0.3783	0.2926
SEGALE	0.3853	0.4353	0.2859
Simple Heuristics			
Token Ratio	-0.0229	0.0211	-0.0013
Sentence Ratio	0.2300	0.2684	0.1783
Sentence Count Difference	-0.2040	-0.2281	-0.1583
Ours (STAR Variants)			
STAR (Default)	0.7524	0.7459	0.7625
w/o SaT (use Spacy)	0.7155	0.6602	0.6132
w/o LaBSE (use M3)	0.7396	0.6621	0.6391

Table 4: Correlation analysis of various metrics with ground-truth structural noise levels.

4 Discussions

4.1 Robust Analysis of STAR

To validate whether STAR accurately reflects structural fidelity, we conduct a sensitivity analysis using a constructed perturbation dataset. Starting from WMT24 (Kocmi et al., 2024a) test set,[8] we introduce synthetic structural noise into target documents by randomly (1) inserting irrelevant sentences (simulating hallucinations), (2) deleting sentences (simulating omissions) , or (3) swapping sentence positions (simulating complex alignments) at ratios varying from 0% to 20%. We then compute Spearman (ρ), Pearson (r) and Kendall (τ) correlations between metric scores and ground-truth noise labels. The results are summarized in Table 4. Furthermore, we conduct two additional analyses: one on a dataset containing *only* insertions and deletions to isolate content errors, and another on the full dataset using a "relaxed" STAR variant where complex alignments are treated as positive matches (numerator) rather than penalties. See Appendix G for detailed results.

Comparison with Existing Alignment Methods. STAR relies on sentence-level alignment as an intermediate step. We benchmark against sentence-level alignment methods like Align-then-Slide (Guo et al., 2025c) and SEGALE (Wang et al., 2025e) by *explicitly* computing STAR from their intermediate alignment outputs. Results show that our Bertalign-based implementation achieves substantially higher correlation with structural errors ($\rho \approx 0.75$ vs. 0.38). This confirms that precise, fine-grained sentence alignment is a prerequisite for reliably measuring structural fidelity in Doc2Doc translation.

Comparison with Length-based Heuristics. We compare STAR with length-based heuristics (Peng et al., 2025a; Guerreiro et al., 2023; Shao et al., 2024a; Hu et al., 2025; Domhan and Zhu, 2025), including *Token Ratio*, *Sentence Ratio*, and *Sentence Count Difference*. As shown in Table 4, these metrics exhibit weak correlations, failing to detect various structural noise when overall document length is preserved. STAR's superior sensitivity confirms that assessing structural fidelity requires verifying sentence-level alignment rather than relying on surface-level statistics.

Comparison to LLM-as-a-Judge Implementation. We further compare STAR with the LLM-as-a-judge version that explicitly assesses document-level structural alignment using Gemini-2.5-Flash. LLM-judge achieves the highest correlation with injected structural noise ($\rho = 0.79$), reflecting its strong semantic reasoning capability. Notably, STAR attains a comparable correlation ($\rho = 0.75$), indicating that it captures structural fidelity in a manner consistent with LLM-based judgments. This suggests that STAR provides a reliable and reproducible alternative signal for structural assessment that is consistent with LLM-based judgments and suitable for practical evaluation settings.

Ablation Study on STAR Components. We evaluate the contribution of key STAR components through ablation experiments: **Segmentation:** Replacing SaT with Spacy (Honnibal et al., 2020) reduces correlation from $\rho \approx 0.75$ to 0.72. This indicates that STAR is relatively robust to sentence boundary conditions. **Encoding:** Switching the embedding model from LaBSE (Feng et al., 2022) to M3 (Chen et al., 2025a) yields a slight decrease in correlation (0.75 vs. 0.74), indicating that STAR is relatively robust to the choice of semantic encoder.

4.2 Comparison with Alternative Optimization Baselines

To thoroughly evaluate the effectiveness of our framework, we compare STAR-masked preference optimization against a diverse set of alternatives, including different data ranking signals and online reinforcement learning (RL) strategies. For

[8] github.com/wmt-conference/wmt24-news-systems

illustration, we use LLaMA-3.1-8B on Zh ⇔ En as a representative. Table 5 shows the results.

Effect of Ranking Metrics on Preference Data Construction. We evaluate the impact of ranking signals by substituting STAR with COMET, COMETKiwi, BLEU, and word-level alignment coverage (Wu et al., 2024c, 2023), under a strictly controlled data budget identical to our main experiment. As shown in Table 5, STAR significantly outperforms all baselines despite the equalized data size. Notably, its superiority over sentence-level methods (He et al., 2024; Agrawal et al., 2024; Tang et al., 2025) confirms that discourse-level structural fidelity is best captured via sentence-to-sentence correspondence.

Comparison with RL We further benchmark our method against standard online RL strategies (GRPO (Shao et al., 2024b), GSPO (Zheng et al., 2025)). Following recent approaches (Feng et al., 2025a,b; He et al., 2025), we directly inject COMET and STAR scores into the reward function for these baselines. As observed in the middle section of Table 5, while online methods utilizing self-generated data generally underperform our offline framework, GSPO paired with STAR achieves competitive results (80.16 COMET). This indicates that while our offline optimization (StarPO) is more effective, STAR nevertheless functions as a robust and high-quality reward signal within RL paradigms. Detailed comparisons with other standard offline preference optimization algorithms (e.g., DPO (Rafailov et al., 2023), SimPO (Meng et al., 2024), ORPO (Hong et al., 2024)) are provided in Appendix H.

Ablation Studies To further dissect the source of our gains, we examine two specific variants in the bottom section of Table 5:

(1) SFT on Preference Data: To isolate the gain attributed to data quality, we fine-tune the model solely on the preferred responses (y_w) from our curated pairs. This baseline yields impressive performance (80.41 and 80.02 COMET), indicating that our STAR-based curation filters for high-quality data. However, our full framework still outperforms this strong baseline (+1.14 and +0.09 COMET), demonstrating that the preference optimization objective effectively leverages the contrastive signal beyond simple supervised imitation.

Method	Zh ⇒ En	En ⇒ Zh
Data Ranking Metrics		
BLEU Ranking	76.55	79.71
COMET Ranking	81.01	79.78
COMETKiwi Ranking	80.56	79.73
Word-level Coverage	75.94	79.46
Online RL Strategies		
GRPO (*w.* COMET)	77.73	79.50
GRPO (*w.* STAR)	77.79	79.44
GRPO (*w.* COMET + STAR)	77.76	79.76
GSPO (*w.* COMET)	77.94	79.77
GSPO (*w.* STAR)	80.16	80.00
GSPO (*w.* COMET + STAR)	77.72	79.97
Our Variants		
SFT on preference data	80.41	80.02
STAR (Relaxed)	80.09	79.59
Random Mask (Sentence level)	<u>81.18</u>	80.05
Random Mask (Token level)	81.17	<u>80.06</u>
CPO	81.10	79.94
Ours (StarPO)	**81.55**	**80.11**

Table 5: COMET scores comparison against alternative optimization strategies. Word-level Coverage is calculated via WSPAlign (Wu et al., 2024c, 2023).

(2) STAR (Relaxed): We treat complex alignments as valid matches. Empirically, this relaxation dilutes the discriminative power, making it difficult to establish a sufficient score margin for pair selection. Consequently, the volume of qualifying training data drops significantly, leading to degraded performance. This confirms that the strict penalization in our standard STAR metric is crucial for generating the sharp preference signals required for effective training, as further illustrated by the score distribution analysis in Appendix I.

(3) Random Masking: We further investigate the effect of loss masking by randomly masking 10% of tokens at both the sentence and token levels. We find that masking serves as an effective regularizer (Gu et al., 2025), outperforming the full-token SFT baseline. This suggests that calculating the loss on a subset of tokens is inherently beneficial; however, our StarPO consistently achieves superior results, demonstrating the advantage of using semantic-aware importance signals over random selection. See Appendix J for detailed analysis of why and how masking works.

4.3 LLM-as-a-judge Metrics Results

Following Sun et al. (2025b), we complement automated metrics with LLM-as-a-judge metrics along multiple orthogonal dimensions, including fluency, content errors, and coherence errors. To avoid self-preference bias (Chen et al., 2025b), we use Gemini-2.5-Flash for all systems. The results

	Fluency ↑	Content ↓	Cohesion ↓
LLAMA-3.1-INSTRUCT			
Base	3.67	2.33	1.95
+ SFT	3.94	1.96	1.70
+ CPO	3.97	1.90	1.58
+ StarPO	3.99	1.40	1.27
QWEN-2.5-7B-INSTRUCT			
Base	3.97	1.64	1.38
+ SFT	4.12	1.63	1.30
+ CPO	4.40	1.39	1.02
+ StarPO	4.45	1.16	0.95
QWEN-3-4B-INSTRUCT			
Base	4.39	1.46	1.13
+ SFT	4.51	1.44	1.11
+ CPO	4.58	1.23	0.91
+ StarPO	4.59	1.17	0.89
OTHER SYSTEMS			
Tower+	4.05	1.29	1.14
GPT-4o	4.18	1.26	1.12
Deepseek-R1	4.37	1.37	1.20

Table 6: LLM-as-a-judge evaluation results. ↑ indicates higher scores are better, while ↓ indicates lower scores are better.

of News-Commentary Zh ⇒ En are shown in Table 6. StarPO consistently achieves superior performance across all dimensions and model families, outperforming both standard CPO and strong baselines like Tower+ and GPT-4o.

5 Related Work

5.1 Document-level Machine Translation

LLM-based document-level machine translation generally falls into two paradigms: Doc2Sent (context-aware), which treats documents as sequences of sentence-level tasks, and Doc2Doc, which processes documents holistically.

In Doc2Sent style, training-free approaches rely on prompting strategies, such as context injection (Wang et al., 2023; Sia and Duh, 2023; Moslem et al., 2023; Zhang et al., 2023a; Lee et al., 2025; Lippmann et al., 2025; Cui et al., 2024; Peng et al., 2025b; Hu et al., 2025), employing self-refinement or memory-based agents (Wang et al., 2025f; Guo et al., 2025b; Koneru et al., 2024). Fine-tuning strategies enhance models via various approaches in data construction (Lyu et al., 2024; Wu et al., 2024a; Zhang et al., 2023b; Stap et al., 2024), and Tower series (Alves et al., 2024; Rei et al., 2024, 2025; Ramos et al., 2025) represent a prominent example. Recent works further analyze the mechanics of context usage (Mąka et al., 2025b,a; Mohammed and Niculae, 2025; Choudhary et al., 2025).

Doc2Doc approaches aim for holistic translation through long-context training (Pang et al., 2025; Li et al., 2026), iterative or agentic refinement (Dong et al., 2025; Li et al., 2025c; Briakou et al., 2024; Wu et al., 2024b), and input optimization strategies like segmentation or knowledge fusion (Hong et al., 2025; Liu et al., 2025). For evaluation, recent works (Guo et al., 2025a,c; Domhan and Zhu, 2025; Wang et al., 2025e; Steingrimsson et al., 2023) predominantly rely on source-target sentence alignment to assess document quality.

5.2 Reinforcement Learning for MT

As references are not necessarily superior to LLM generations, RL (Ouyang et al., 2022) becomes essential for advancing MT. Recent research focus on training specialized reward models (Li et al., 2025a; Feng et al., 2025c; Ramos et al., 2024; Tan and Monz, 2025) to guide this process.

RL approaches are generally categorized into online and offline methods. Online methods and reward-based methods, commonly use Quality Estimation models as reward models. (He et al., 2024). Such online frameworks is also compatible for training large reasoning models (He et al., 2025; Feng et al., 2025a,b; Wang et al., 2025b,c,d; Li et al., 2025b). Typically, these methods apply various items into reward models. Conversely, offline methods like DPO (Rafailov et al., 2023) and its variants (Ethayarajh et al., 2024; Meng et al., 2024; Xu et al., 2024, 2025; Zeng et al., 2024) rely on pre-curated preference datasets for stability. Various metrics and approaches are proposed to construct and leverage these datasets. (Sun et al., 2025a; Cui et al., 2025; Wu et al., 2024c; Tang et al., 2025; Zhong et al., 2025; Agrawal et al., 2024; Wang et al., 2025a; Yang et al., 2024).

6 Conclusion

To address structural misalignment in Doc2Doc translation, we introduce **STAR**, a metric for evaluating document-level structural fidelity, and **StarPO**, a preference optimization framework that utilizes dynamic masking to target omissions and hallucinations. Experiments demonstrate that StarPO enables compact models to surpass massive proprietary systems like GPT-4o in translation quality while significantly improving token efficiency. This work establishes a robust paradigm for faithful, "in-one-go" long-document translation without complex agentic workflows.

Limitations

First, in "in-one-go" Doc2Doc scenarios, establishing sentence-level alignment is an unavoidable prerequisite for calculating any fine-grained quality metric (e.g., d-COMET). Second, while enforcing 1-to-1 alignment effectively mitigates hallucinations, it imposes a structural rigidity that could theoretically discourage valid complex mappings in stylized texts, though our empirical results suggest this impact is minimal. Third, our experimental validation is currently concentrated on compact models (4B to 9B parameters) and high-to-medium resource languages. Validating the scalability of StarPO to larger architectures (e.g., 70B+) and low-resource languages remains a critical direction for future research. Finally, regarding data construction, we currently leverage proprietary models (e.g., GPT-4o) to augment candidate diversity. Although the selection of high-quality samples is strictly governed by our own STAR metric, this reliance on commercial APIs for initial generation currently prevents a fully end-to-end open-source pipeline. Future work aims to substitute this step with open-source alternatives, thereby enabling a completely offline-deployable training framework.

References

Sweta Agrawal, José G. C. De Souza, Ricardo Rei, António Farinhas, Gonçalo Faria, Patrick Fernandes, Nuno M Guerreiro, and Andre Martins. 2024. Modeling user preferences with automatic metrics: Creating a high-quality preference dataset for machine translation. In *Proceedings of EMNLP*, pages 14503–14519.

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. *CoRR*, abs/2402.17733.

Eleftheria Briakou, Jiaming Luo, Colin Cherry, and Markus Freitag. 2024. Translating step-by-step: Decomposing the translation process for improved translation quality of long-form texts. In *Proceedings of WMT*, pages 1301–1317.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2025a. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *CoRR*, abs/2402.03216.

Zhi-Yuan Chen, Hao Wang, Xinyu Zhang, Enrui Hu, and Yankai Lin. 2025b. Beyond the surface: Measuring self-preference in LLM judgments. In *Proceedings of EMNLP*, pages 1653–1672, Suzhou, China.

Ritvik Choudhary, Rem Hida, Masaki Hamada, Hayato Futami, and Toshiyuki Sekiya. 2025. Exploring context strategies in LLMs for discourse-aware machine translation. In *Findings of EMNLP*, pages 24382–24391.

Guofeng Cui, Pichao Wang, Yang Liu, Zemian Ke, Zhu Liu, and Vimal Bhat. 2025. CRPO: Confidence-reward driven preference optimization for machine translation. In *Findings of ACL*, pages 560–574.

Menglong Cui, Jiangcun Du, Shaolin Zhu, and Deyi Xiong. 2024. Efficiently exploring large language models for document-level machine translation with in-context learning. In *Findings of ACL*, pages 10885–10897.

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

Tobias Domhan and Dawei Zhu. 2025. Same evaluation, more tokens: On the effect of input length for machine translation evaluation using large language models. In *Proceedings of EMNLP*, pages 7940–7958.

Yichen Dong, Xinglin Lyu, Junhui Li, Daimeng Wei, Min Zhang, Shimin Tao, and Hao Yang. 2025. Two intermediate translations are better than one: Fine-tuning LLMs for document-level translation refinement. In *Proceedings of ACL*, pages 14917–14933.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *CoRR*, abs/2402.01306.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of ACL*, pages 878–891.

Zhaopeng Feng, Shaosheng Cao, Jiahan Ren, Jiayuan Su, Ruizhe Chen, Yan Zhang, Jian Wu, and Zuozhu Liu. 2025a. MT-r1-zero: Advancing LLM-based machine translation via r1-zero-like reinforcement learning. In *Findings of the EMNLP*, pages 18685–18702.

Zhaopeng Feng, Yupu Liang, Shaosheng Cao, Jiayuan Su, Jiahan Ren, Zhe Xu, Yao Hu, Wenxuan Huang, Jian Wu, and Zuozhu Liu. 2025b. Mt^3: Scaling mllm-based text image machine translation via multi-task reinforcement learning.

Zhaopeng Feng, Jiahan Ren, Jiayuan Su, Jiamei Zheng, Hongwei Wang, and Zuozhu Liu. 2025c. MT-RewardTree: A comprehensive framework for advancing LLM-based machine translation via reward

modeling. In *Findings of EMNLP*, pages 18556–18567.

Markus Frohmann, Igor Sterner, Ivan Vulić, Benjamin Minixhofer, and Markus Schedl. 2024. Segment any text: A universal approach for robust, efficient and adaptable sentence segmentation. In *Proceedings of EMNLP*, pages 11908–11941.

Yuzhe Gu, Wenwei Zhang, Chengqi Lyu, Dahua Lin, and Kai Chen. 2025. Mask-DPO: Generalizable fine-grained factuality alignment of LLMs. In *The Thirteenth International Conference on Learning Representations*.

Nuno M. Guerreiro, Elena Voita, and André Martins. 2023. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. In *Proceedings of EACL*, pages 1059–1075. Association for Computational Linguistics.

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

Jiaxin Guo, Yuanchang Luo, Daimeng Wei, Ling Zhang, Zongyao Li, Hengchao Shang, Zhiqiang Rao, Shaojun Li, Jinlong Yang, Zhanglin Wu, and Hao Yang. 2025b. Doc-guided sent2sent++: A sent2sent++ agent with doc-guided memory for document-level machine translation. *CoRR*, abs/2501.08523.

Jiaxin Guo, Daimeng Wei, Yuanchang Luo, Xiaoyu Chen, Zhanglin Wu, Huan Yang, Hengchao Shang, Zongyao Li, Zhiqiang Rao, Jinlong Yang, and Hao Yang. 2025c. Align-then-slide: A complete evaluation framework for ultra-long document-level machine translation. *CoRR*, abs/2509.03809.

Minggui He, Yilun Liu, Shimin Tao, Yuanchang Luo, Hongyong Zeng, Chang Su, Li Zhang, Hongxia Ma, Daimeng Wei, Weibin Meng, and 1 others. 2025. R1-t1: Fully incentivizing translation capability in llms via reasoning learning. *CoRR*, abs/2502.19735.

Zhiwei He, Xing Wang, Wenxiang Jiao, Zhuosheng Zhang, Rui Wang, Shuming Shi, and Zhaopeng Tu. 2024. Improving machine translation with human feedback: An exploration of quality estimation as a reward model. In *Proceedings of NAACL:HLT*, pages 8164–8180. Association for Computational Linguistics.

Hanghai Hong, Yibo Xie, Jiawei Zheng, and Xiaoli Wang. 2025. SubDocTrans: Enhancing document-level machine translation with plug-and-play multi-granularity knowledge augmentation. In *Findings of EMNLP*, pages 14490–14506.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. Orpo: Monolithic preference optimization without reference model. In *Proceedings of EMNLP*, pages 11170–11189.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd, and 1 others. 2020. spacy: Industrial-strength natural language processing in python.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. In *Proceedings of ICLR*.

Hanxu Hu, Jannis Vamvas, and Rico Sennrich. 2025. Source-primed multi-turn conversation helps large language models translate documents. In *Findings of EMNLP*, pages 23702–23712.

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. In *Proceedings of WMT*, pages 419–451.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondrej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popovic, Mariya Shmatova, and 2 others. 2024a. Preliminary wmt24 ranking of general mt systems and llms. *CoRR*, abs/2407.19884.

Tom Kocmi, Vilém Zouhar, Christian Federmann, and Matt Post. 2024b. Navigating the metrics maze: Reconciling score magnitudes and accuracies. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1999–2014, Bangkok, Thailand.

Sai Koneru, Miriam Exel, Matthias Huck, and Jan Niehues. 2024. Contextual refinement of translations: Large language models for sentence and document-level post-editing. In *Proceedings of NAACL:HLT*, pages 2711–2725.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Minjae Lee, Youngbin Noh, and Seung Jin Lee. 2025. A testset for context-aware LLM translation in Korean-to-English discourse level translation. In *Proceedings of COLING*, pages 1632–1646.

Tianjiao Li, Mengran Yu, Chenyu Shi, Yanjun Zhao, Xiaojing Liu, Qi Zhang, Xuanjing Huang, Qiang Zhang, and Jiayin Wang. 2025a. RIVAL: Reinforcement learning with iterative and adversarial optimization for machine translation. In *Findings of EMNLP 2025*, pages 3064–3079.

Yachao Li, Junhui Li, Jing Jiang, and Min Zhang. 2026. Enhancing document-level translation of

large language model via translation mixed instructions. *Neurocomputing*, 664:132041.

Zihao Li, Shaoxiong Ji, and Jörg Tiedemann. 2025b. Test-time scaling of reasoning models for machine translation. *CoRR*, abs/2510.06471.

Zongyao Li, Zhiqiang Rao, Hengchao Shang, Jiaxin Guo, Shaojun Li, Daimeng Wei, and Hao Yang. 2025c. Enhancing large language models for document-level translation post-editing using monolingual data. In *Proceedings of COLING*, pages 8830–8840.

Philip Lippmann, Konrad Skublicki, Joshua Tanner, Shonosuke Ishiwatari, and Jie Yang. 2025. Context-informed machine translation of manga using multimodal large language models. In *Proceedings of COLING*, pages 3444–3464.

Bin Liu, Xinglin Lyu, Junhui Li, Daimeng Wei, Min Zhang, Shimin Tao, and Hao Yang. 2025. Improving llm-based document-level machine translation with multi-knowledge fusion. *CoRR*, abs/2503.12152.

Lei Liu and Min Zhu. 2023. Bertalign: Improved word embedding-based sentence alignment for chinese–english parallel corpora of literary texts. *Digital Scholarship in the Humanities*, 38:621–634.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Xinglin Lyu, Junhui Li, Yanqing Zhao, Min Zhang, Daimeng Wei, Shimin Tao, Hao Yang, and Min Zhang. 2024. DeMPT: Decoding-enhanced multi-phase prompt tuning for making LLMs be better context-aware translators. In *Proceedings of EMNLP*, pages 20280–20295.

Paweł Mąka, Yusuf Can Semerci, Jan Scholtes, and Gerasimos Spanakis. 2025a. Analyzing the attention heads for pronoun disambiguation in context-aware machine translation models. In *Proceedings of COLING*, pages 6348–6377.

Paweł Mąka, Yusuf Can Semerci, Jan Scholtes, and Gerasimos Spanakis. 2025b. You are what you train: Effects of data composition on training context-aware machine translation models. In *Proceedings of EMNLP*, pages 27402–27425.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *NIPS*, 37:124198–124235.

Wafaa Mohammed and Vlad Niculae. 2025. Context-aware or context-insensitive? assessing LLMs' performance in document-level translation. In *Proceedings of Machine Translation Summit XX: Volume 1*, pages 126–137.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models. In *Proceedings of EAMT*, pages 227–237.

OpenAI. 2024. Gpt-4o mini: advancing cost-efficient intelligence. https://openai.com/research/gpt-4.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *NIPS*, 35:27730–27744.

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

Ziqian Peng, Rachel Bawden, and François Yvon. 2025a. Investigating length issues in document-level machine translation. In *Proceedings of Machine Translation Summit XX: Volume 1*, pages 4–23, Geneva, Switzerland.

Ziqian Peng, Rachel Bawden, and François Yvon. 2025b. Self-retrieval from distant contexts for document-level machine translation. In *Proceedings of WMT*, pages 220–240, Suzhou, China.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Proceedings of NIPS*, pages 53728–53741.

Miguel Moura Ramos, Tomás Almeida, Daniel Vareta, Filipe Azevedo, Sweta Agrawal, Patrick Fernandes, and André F. T. Martins. 2024. Fine-grained reward optimization for machine translation using error severity mappings. *CoRR*, abs/2411.05986.

Miguel Moura Ramos, Patrick Fernandes, Sweta Agrawal, and André F. T. Martins. 2025. Multilingual contextualization of large language models for document-level machine translation. *CoRR*, abs/2504.12140.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of WMT*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid).

Ricardo Rei, Nuno M. Guerreiro, José Pombal, João Alves, Pedro Teixeirinha, Amin Farajian, and André F. T. Martins. 2025. Tower+: Bridging generality and translation specialization in multilingual llms. *CoRR*, abs/2506.17080.

Ricardo Rei, Jose Pombal, Nuno M. Guerreiro, João Alves, Pedro Henrique Martins, Patrick Fernandes, Helena Wu, Tania Vaz, Duarte Alves, Amin Farajian, Sweta Agrawal, Antonio Farinhas, José G. C. De Souza, and André Martins. 2024. Tower v2: Unbabel-IST 2024 submission for the general MT shared task. In *Proceedings of WMT*, pages 185–204.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C. Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte M. Alves, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022b. Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task. In *Proceedings WMT*, pages 634–645.

Chenze Shao, Fandong Meng, Jiali Zeng, and Jie Zhou. 2024a. Understanding and addressing the under-translation problem from the perspective of decoding objective. In *Proceedings of ACL*, pages 3800–3814.

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

Ibne Farabi Shihab, Sanjeda Akter, and Anuj Sharma. 2025. Detecting and mitigating reward hacking in reinforcement learning systems: A comprehensive empirical study. *CoRR*, abs/2507.05619.

Suzanna Sia and Kevin Duh. 2023. In-context learning as maintaining coherency: A study of on-the-fly machine translation using large language models. In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 173–185.

David Stap, Eva Hasler, Bill Byrne, Christof Monz, and Ke Tran. 2024. The fine-tuning paradox: Boosting translation quality without sacrificing LLM abilities. In *Proceedings of ACL*, pages 6189–6206.

Steinthor Steingrimsson, Hrafn Loftsson, and Andy Way. 2023. SentAlign: Accurate and scalable sentence alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 256–263, Singapore. Association for Computational Linguistics.

Haoxiang Sun, Ruize Gao, Pei Zhang, Baosong Yang, and Rui Wang. 2025a. Enhancing machine translation with self-supervised preference data. In *Proceedings of ACL*, pages 23916–23934.

Yirong Sun, Dawei Zhu, Yanjun Chen, Erjia Xiao, Xinghao Chen, and Xiaoyu Shen. 2025b. Fine-grained and multi-dimensional metrics for document-level machine translation. In *Proceedings of NAACL:HLT*, pages 1–17.

Shaomu Tan and Christof Monz. 2025. ReMedy: Learning machine translation evaluation from human preferences with reward modeling. In *Proceedings of EMNLP*, pages 4370–4387.

Zilu Tang, Rajen Chatterjee, and Sarthak Garg. 2025. Mitigating hallucinated translations in large language models with hallucination-focused preference optimization. In *Proceedings of NAACL:HLT*, pages 3410–3433.

Llama Team. 2024a. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Qwen Team. 2024b. Qwen2.5: A party of foundation models.

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 WMT*, pages 118–128.

Hao Wang, Linlong Xu, Heng Liu, Yangyang Liu, Xiaohu Zhao, Bo Zeng, Liangying Shao, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2025a. Beyond single-reward: Multi-pair, multi-perspective preference optimization for machine translation. *CoRR*, abs/2510.13434.

Jiaan Wang, Fandong Meng, Yunlong Liang, and Jie Zhou. 2025b. DRT: Deep reasoning translation via long chain-of-thought. In *Findings of ACL 2025*, pages 6770–6782.

Jiaan Wang, Fandong Meng, and Jie Zhou. 2025c. Deeptrans: Deep reasoning translation via reinforcement learning.

Jiaan Wang, Fandong Meng, and Jie Zhou. 2025d. Extrans: Multilingual deep reasoning translation via exemplar-enhanced reinforcement learning. *CoRR*, abs/2505.12996.

Kuang-Da Wang, Shuoyang Ding, Chao-Han Huck Yang, Ping-Chun Hsieh, Wen-Chih Peng, Vitaly Lavrukhin, and Boris Ginsburg. 2025e. Extending automatic machine translation evaluation to book-length documents. In *Proceedings of EMNLP*, pages 32311–32327.

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

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 EMNLP*, pages 16646–16661.

Yutong Wang, Jiali Zeng, Xuebo Liu, Derek F. Wong, Fandong Meng, Jie Zhou, and Min Zhang. 2025f. Delta: An online document-level translation agent based on multi-level memory. In *Proceedings of ICLR*.

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George Foster, and Gholamreza Haffari. 2024a. Adapting large language models for document-level machine translation. *CoRR*, abs/2401.06468.

Minghao Wu, Jiahao Xu, and Longyue Wang. 2024b. TransAgents: Build your translation company with language agents. In *Proceedings of EMNLP: System Demonstrations*, pages 131–141.

Qiyu Wu, Masaaki Nagata, Zhongtao Miao, and Yoshimasa Tsuruoka. 2024c. Word alignment as preference for machine translation. In *Proceedings of EMNLP*, pages 3223–3239.

Qiyu Wu, Masaaki Nagata, and Yoshimasa Tsuruoka. 2023. WSPAlign: Word alignment pre-training via large-scale weakly supervised span prediction. In *Proceedings of ACL*, pages 11084–11099.

Haoran Xu, Kenton Murray, Philipp Koehn, Hieu Hoang, Akiko Eriguchi, and Huda Khayrallah. 2025. X-ALMA: Plug & play modules and adaptive rejection for quality translation at scale. In *Proceedings of ICLR*.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of LLM performance in machine translation. In *Proceedings of ICML*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 40 others. 2025. Qwen3 technical report. *CoRR*, abs/2505.09388.

Guangyu Yang, Jinghong Chen, Weizhe Lin, and Bill Byrne. 2024. Direct preference optimization for neural machine translation with minimum Bayes risk decoding. In *Proceedings of NAACL:HLT (Volume 2: Short Papers)*, pages 391–398, Mexico City, Mexico.

Armel Randy Zebaze, Benoît Sagot, and Rachel Bawden. 2025. In-context example selection via similarity search improves low-resource machine translation. In *Findings of NAACL*, pages 1222–1252.

Jiali Zeng, Fandong Meng, Yongjing Yin, and Jie Zhou. 2024. Teaching large language models to translate with comparison. In *Proceedings of AAAI*, pages 19488–19496.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023a. Prompting large language model for machine translation: A case study. In *ICML*, pages 41092–41110.

Xuan Zhang, Navid Rajabi, Kevin Duh, and Philipp Koehn. 2023b. Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with QLoRA. In *Proceedings of WMT*, pages 468–481.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. 2025. Swift:a scalable lightweight infrastructure for fine-tuning. In *Proceedings of AAAI*, pages 29733–29735.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. 2025. Group sequence policy optimization. *CoRR*, abs/2507.18071.

Zhiqiang Zhong, Simon Sataa-Yu Larsen, Haoyu Guo, Tao Tang, Kuangyu Zhou, and Davide Mottin. 2025. Automatic annotation augmentation boosts translation between molecules and natural language. In *Findings of NAACL*, pages 6177–6194.

Vilém Zouhar, Pinzhen Chen, Tsz Kin Lam, Nikita Moghe, and Barry Haddow. 2024. Pitfalls and outlooks in using COMET. In *Proceedings of WMT*, pages 1272–1288.

A Efficiency Analyses

Figure 2 compares the alignment score vs. token consumption among several typical Doc2Sent and Doc2Doc systems.

- **Doc2Sent (w=3):** Adopts a standard sliding context window strategy, similar to Wu et al. (2024a); Lyu et al. (2024); Koneru et al. (2024); Cui et al. (2024) where the current sentence is translated by conditioning on the source and target history of the preceding 3 sentences (i.e. w=3).

- **Doc2Sent (Source-primed):** Models document translation as a sequential multi-turn dialogue, requiring the model to translate sentence-by-sentence while maintaining the full conversation history. (Hu et al., 2025)

- **Doc2Sent (DelTA):** An agentic framework that leverages autonomous agents to handle context-aware sentence translation (Wang et al., 2025f).

- **Doc2Doc (KFMT):** Incorporates auxiliary knowledge via multi-turn interactions prior to generating the full document (Liu et al., 2025). We evaluate two variants: one using explicit **sentence delimiters** to structure the output, and one generating raw text without delimiters.

- **Doc2Doc (DocRefine):** A multi-stage translation refinement approach that first generates two drafts and then iteratively improves the document's coherence and accuracy (Dong et al., 2025).

- **Doc2Doc (Mix-level SFT):** We use trained models by MixSFT (Li et al., 2026). Similar to KFMT, this is evaluated both **with and without sentence delimiters**.

- **Doc2Doc (DeepSeek-R1):** Represents state-of-the-art Large Reasoning Models (LRMs), leveraging intrinsic chain-of-thought reasoning capabilities for complex translation tasks.

- **Doc2Doc(GPT-4o):** Serves as a strong proprietary baseline, queried using standard document-level translation prompts without additional fine-tuning.

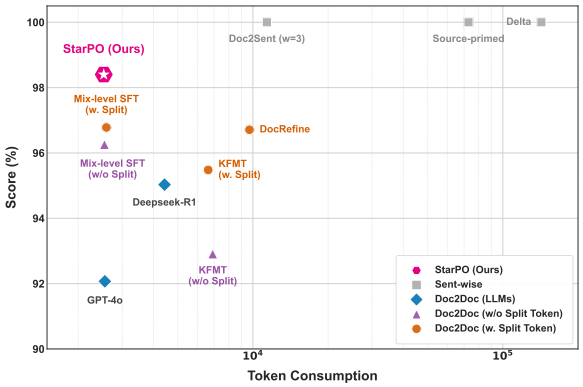

Figure 2: **Alignment score vs. Token Consumption.** We compare our proposed **StarPO** against various baselines, including sentence-level systems, document-level baselines, and proprietary LLMs (e.g., GPT-4o, Deepseek-R1). Note that the x-axis is plotted on a logarithmic scale (10^n).

- **Doc2Doc(StarPO):** Our proposed framework that optimizes the model using preference pairs curated via the STAR metric, specifically designed to enforce strict structural alignment alongside semantic fidelity.

B Detailed Case Study

While the STAR metric promotes strict 1-to-1 sentence alignment, it is crucial to understand why deviations from this pattern typically indicate degradation in translation quality. In this section, we categorize common alignment pathologies and analyze their negative impact on fidelity, readability, and information completeness. As summarized in Figure 3, each non-1-to-1 alignment type corresponds to specific translation errors:

- **k-to-1 (Information Loss via Merging):** While merging sentences can sometimes be a stylistic choice, automated systems frequently use it to compress information, resulting in the loss of nuanced details or the omission of secondary clauses from the source.

- **1-to-k (Distortion via Splitting):** Splitting a single coherent sentence into multiple fragments often disrupts the discourse flow. It can not only introduce redundant subjects or unnatural pauses but also sometimes leads to information loss, where subordinate clauses or subtle nuances are inadvertently omitted during fragmentation, where subordinate clauses or subtle nuances are inadvertently omitted during fragmentation.

Align type	Source Text	Target Text
1-to-0	The proposal was accepted unanimously. However, concerns regarding the budget remain.	提案全票通过。
0-to-1	The weather forecast predicts rain tomorrow.	天气预报显示明天有雨。 建议居民待在室内。
k-to-1	The CEO finished his speech and stepped off the podium. The audience burst into applause.	CEO结束演讲时，观众爆发出了热烈的掌声。
1-to-k	This new pattern - which persisted through the 1908s and 1990s, and accelerated after 2000 - caused inequality to rise sharply, weakening the foundations of globalization.	这个新的模式一直持续到 80 年代 90 年代，甚至加速了。 在此期间，贫富差距急剧扩大，这削弱了全球化根基。
k-to-k'	But, in my view, Trump seems to be sending a deeper message about corporate decision-making. Despite Trump's own business record – which, his opponents will point out, includes multiple bankruptcies and non-payment of contractors and their workers – it's possible that he is now trying to change a business and investment culture that elevates the interests of capital, corporations, and shareholders, and treats labor as expendable.	但我的看法是，特朗普希望传递一个更深层次的信息：资本、公司和股东的利益比工人的价值更重要。 考虑到特朗普自己的商业记录，包括教宗破产案以及欠薪拖欠承包商和他们的工人——他的政敌会争辩说，所以这是一个颠覆过去两百年来的商业和投资思潮的时候了。

Figure 3: Detailed case study illustrating alignment challenges: (1) 1-to-0 alignments result in obvious source omissions; (2) 0-to-1 alignments contain hallucinations not present in the original text; (3) k-to-1 alignments exhibit lowered readability; (4) 1-to-k alignments risk semantic loss; and (5) k-to-k' alignments scramble sentence order, impeding source correspondence. Corresponding texts in the source and target are highlighted in red and blue, respectively.

- **k-to-k' (Structural Chaos):** Complex many-to-many alignments often indicate a breakdown in the linear correspondence of the document. Although the semantic information might be partially preserved, the reordering of logic and the scrambling of sentence boundaries significantly reduce readability and increase the cognitive load for the reader.

C LLM-judge version of STAR

To validate the accuracy of our automated STAR metric, we implement an LLM-based version using the prompt template illustrated in Figure 4. This prompt is designed to simulate human-level judgment on document structure.

D Data Statistics

In this section, we provide detailed statistics regarding the bilingual parallel corpora utilized in our experiments: the News-Commentary dataset and the Guofeng dataset. The statistical summaries for these datasets are presented in Table 7 and Table 8, respectively.

The News-Commentary dataset encompasses a diverse range of language pairs, including Chinese-English (ZH⇔EN), Chinese-German (ZH⇔DE), Russian-English (RU⇔EN), German-English (DE⇔EN), and Spanish-English (ES⇔EN). The Guofeng dataset primarily focuses on pairs involving Chinese, specifically Chinese-Russian (ZH⇔RU), Chinese-English (ZH⇔EN), and Chinese-German (ZH⇔DE).

Dataset	#Document Train/Valid/Test	#Document for StarPO	Average Tokens	Max Tokens
De ⇒ En	8.4K/150/150	1,008	1,797	6,540
En ⇒ De	8.4K/150/150	734	1,066	4,065
Es ⇒ En	9.7K/150/150	3,325	1,643	6,293
En ⇒ Es	9.7K/150/150	3,937	1,071	4,146
Ru ⇒ En	7.3K/150/150	1,744	1,776	7,951
En ⇒ Ru	7.3K/150/150	1,491	1,079	5,557
Zh ⇒ En	8.6K/150/150	1,253	1,377	4,425
En ⇒ Zh	8.6K/150/150	1,753	1,090	3,609
Zh ⇒ De	7.7K/150/150	2,504	1,357	5,912
De ⇒ Zh	7.7K/150/150	1,364	1,828	7,215

Table 7: Statistics of the News-Commentary dataset.

For each dataset and language direction, we report four key metrics based on document-level analysis: the quantity of documents across the training, validation, and test splits; the size of the dataset used for preference optimization; the average number of tokens per document; and the maximum number of tokens found in a single document within the corpus.

E Implementation Details

We implement our experiments in ms-swift (Zhao et al., 2025)[9]. During fine-tuning, we adapt LoRA (Hu et al., 2021). We set LoRA rank to 8 and LoRA alpha to 16 , respectively. The models

[9] https://github.com/modelscope/ms-swift

Prompts for LLM-judging STAR

You are an expert in Bitext Alignment and Translation Quality Assessment.
Please split the source and target documents into sentences, align the source sentences and target sentences and calculate the Sentence Translation Alignment Rate.

- Task: Segment the texts into aligned groups/units.
- Classify Units: Classify each unit into one of the following categories:
  - 1-to-1 (Strict Match): 1 Source sentence aligns exactly to 1 Target sentence.
  - 1-to-k (Split): 1 Source sentence is split into N Target sentences (k>1).
  - k-to-1 (Merge): k Source sentences are merged into 1 Target sentence (k>1).
  - k-to-k′ (Cross/Complex): k Source sentences align to k′ Target sentences as a block (k>1, k′>1).
  - 1-to-0 (Omission): 1 Source sentence has no corresponding translation.
  - 0-to-1 (Hallucination): 1 Target sentence has no corresponding source sentence.
- Calculation Formula: Count the number of Alignment Units (relationships), not sentences.
  For example, a "1-to-3" split counts as 1 unit.
  $$ Score = \frac{Count_{1:1}}{Count_{Total}} $$
  where $Count_{Total}$ is the sum of the counts of all 6 types listed above.
- Output Format:
  First, provide a "Reasoning Step" listing any non-1-to-1 alignments found (e.g., "Source [5] -> Target [5,6] (Split)").
  Then, strictly output the metrics in JSON format:
  ```json
{"count_1_to_1": <int>, "count_1_to_n": <int>, "count_n_to_1": <int>, "count_n_to_m": <int>, "count_1_to_0": <int>, "count_0_to_1": <int>, "total_units": <int>, "alignment_rate": <float>}
  ```

- Input Data:
  - Source Text:
    ```
    <src_doc>
    ```
  - Target Text:
    ```
    <tgt_doc>
    ```

Figure 4: The prompt template used for the LLM-as-a-judge implementation of STAR. The final score is calculated strictly based on the ratio of 1-to-1 matches to total alignment units, serving as a high-precision reference for validating our automated metric.

Dataset	#Document Train/Valid/Test	#Document for StarPO	Average Tokens	Max Tokens
Zh ⇒ En	22.0K/25/25	2.0K	1,853	12,961
En ⇒ Zh	22.0K/25/25	2.0K	1,624	10,956
Zh ⇒ De	6.0K/30/30	2.0K	1,927	7,962
De ⇒ Zh	6.0K/30/30	2.0K	2,392	9,909
Zh ⇒ Ru	6.0K/30/30	2.0K	1,991	6,134
Ru ⇒ Zh	6.0K/30/30	2.0K	2,470	7,514

Table 8: Statistics of the Guofeng dataset.

are trained for 1 epoch using AdamW optimizer with learning rate of 1×10^{-4}, warmup ratio of 0.05. We set β to 0.1. Our experiments run on one NVIDIA H100 GPU, requiring approximately 1 hour for training. During inferencing, we set temperature to 0.3, beam size to 1 in vllm (Kwon et al., 2023) framework. The specific prompt template used in our experiments is illustrated in Figure 5.

F Detailed Metric Scores

We report d-COMETKiwi scores (via `wmt22-cometkiwi-da` (Rei et al., 2022b)[10]) for the News-Commentary and Guofeng test sets in Tables 9 and 15. The structural fidelity of these systems, as measured by our proposed STAR score, is detailed in Tables 10 and 14. Additionally, we report the alignment results on the News-Commentary Zh ⇒ En dataset

[10] https://huggingface.co/Unbabel/wmt22-cometkiwi-da

System	Zh ⇔ En		De ⇔ En		De ⇔ Zh		Ru ⇔ En		En ⇔ Es		Avg.
	⇒	⇐	⇒	⇐	⇒	⇐	⇒	⇐	⇒	⇐	
LLAMA-3.1-8B-INSTRUCT											
Base	68.46	72.61	79.54	81.77	71.99	71.58	82.65	79.50	84.09	81.95	77.41
+ SFT	74.86	77.46	82.25	83.03	70.79	70.95	82.62	79.55	81.97	83.69	78.72
+ CPO	75.24	75.19	82.64	83.63	72.01	72.10	82.70	81.29	84.41	84.40	79.36
+StarPO	77.15	78.49	**82.76**	**83.98**	72.60	**74.90**	82.67	81.37	**84.69**	**84.70**	80.33
QWEN2.5-7B-INSTRUCT											
Base	73.78	73.33	79.66	80.93	74.29	69.01	80.99	77.71	78.29	81.10	76.91
+SFT	76.02	74.02	80.27	81.30	74.88	71.46	83.12	81.43	81.33	82.05	78.59
+CPO	77.90	79.08	82.04	82.14	75.70	72.29	83.04	81.75	83.61	84.20	80.09
+StarPO	78.35	79.21	82.67	82.26	**75.99**	72.51	**83.39**	83.05	84.15	84.66	80.62
QWEN3-4B-INSTRUCT											
Base	80.34	83.55	82.49	82.98	74.26	69.50	82.42	84.09	83.77	84.57	80.79
+SFT	80.53	83.99	82.48	83.19	74.12	69.96	82.86	84.15	83.86	84.63	80.98
+CPO	80.68	83.91	82.44	83.20	74.89	70.13	82.89	83.95	83.42	84.68	81.02
+StarPO	**81.18**	**84.14**	82.63	83.86	75.71	71.48	83.21	**84.30**	84.41	**84.70**	**81.56**
OTHER SYSTEMS											
Tower+	77.31	77.03	82.21	82.78	75.06	71.57	83.30	82.95	84.60	84.16	80.10
GPT-4o	76.92	78.40	82.35	81.79	75.38	71.91	82.93	83.46	83.02	84.33	80.05
Deepseek-R1	79.14	81.54	80.71	82.38	75.33	70.95	81.17	83.79	83.55	82.29	80.09

Table 9: Performance in dCOMETKiwi scores on the News-Commentary test set. **Bold** scores represent the global best performance and underlined scores represent the global second-best performance. Blue text background indicates that the improvement over the origin Base model achieves at least 85% accuracy with the human judgment (Kocmi et al., 2024b). Specifically, the improvement needs a minimin of ≥ 0.67 for `wmt22-cometkiwi-da`.

Prompts for In-one-go Document-level Translation

Translate this document into <*tgt_lang*> without any explanations.

```
<src_doc>
```

Figure 5: **The universal prompt template used for document-level translation.** To ensure a fair comparison and eliminate prompt engineering variance, we apply this standardized Doc2Doc instruction across all experiments. <*tgt_lang*> represents the target language, and <*src_doc*> represents the source document, respectively.

evaluated by Gemini-2.5-Flash using the prompt from Appendix C, as shown in Table 11.

G Detailed Robust Analyses of STAR

Performance under Relax Constraints. To further assess robustness, we conduct a sensitivity analysis using a Relax evaluation protocol. Unlike the strict 1-to-1 requirement used for training data selection, this setting treats all aligned contentincluding complex mappings (k-to-k', 1-to-k, k-to-1)as valid positive matches:

$$\text{STAR}_{\text{relax}}(S,T) = \frac{|\mathcal{U}_{1:1}| + |\mathcal{U}_{\text{complex}}|}{|\mathcal{U}_{1:1}| + |\mathcal{U}_{1:0}| + |\mathcal{U}_{0:1}| + |\mathcal{U}_{\text{complex}}|}. \tag{6}$$

Consequently, metrics are penalized exclusively for completely unaligned segments, corresponding to severe errors like hallucinations (0-to-1) or omissions (1-to-0).

To ensure a fair comparison, we applied this same relaxation logic to all alignment-based baselines, including Guo et al. (2025c); Wang et al. (2025e). As shown in Table 12, even under these looser constraints, STAR demonstrates superior sensitivity compared to other alignment methods and length-based heuristics. Notably, it achieves correlations ($\rho = 0.7449$) comparable to the LLM-judge ($\rho = 0.7766$), confirming that STAR accurately penalizes critical semantic errors while remaining robust to acceptable structural shifts.

Sensitivity to Insertions and Deletions (No Complex Alignments). To isolate the metric's sensitivity to critical content errors without the interference of structural reordering, we constructed a "Simplified Structural Noise" dataset. In this setting, we strictly excluded complex alignment scenarios (e.g., n-to-m merges, splits, or sentence swaps). The dataset consists exclusively of 1-to-1 matches interspersed with varying ratios of pure

System	Zh ⇔ En		De ⇔ En		De ⇔ Zh		Ru ⇔ En		En ⇔ Es		Avg.
	⇒	⇐	⇒	⇐	⇒	⇐	⇒	⇐	⇒	⇐	
LLAMA-3.1-8B-INSTRUCT											
Base	90.31	87.21	92.52	95.41	78.96	82.79	97.05	93.92	90.97	94.06	90.32
+ SFT	93.52	91.75	95.77	95.60	80.39	93.58	96.97	97.58	77.13	95.64	91.79
+ CPO	95.00	91.72	95.99	95.52	79.83	90.41	97.30	97.65	90.64	97.50	93.16
+StarPO	95.62	91.09	96.80	95.59	80.80	97.54	97.27	**98.01**	90.13	**97.63**	94.05
QWEN2.5-7B-INSTRUCT											
Base	94.97	93.92	95.85	95.10	61.80	92.03	98.05	86.87	96.02	97.33	91.19
+SFT	95.51	92.96	94.54	94.45	88.42	93.25	97.49	95.55	81.10	97.32	93.06
+CPO	95.40	93.36	95.89	95.36	95.92	92.49	97.64	95.54	92.83	97.56	95.20
+StarPO	**96.07**	95.67	96.55	95.36	**96.83**	93.12	**98.15**	97.71	96.21	97.37	**96.30**
QWEN3-4B-INSTRUCT											
Base	94.67	93.41	96.33	96.45	82.66	91.01	97.69	96.92	96.12	91.38	93.66
+SFT	95.02	92.92	96.48	91.05	83.03	91.26	97.48	95.35	95.36	96.12	93.41
+CPO	95.01	93.24	96.60	95.33	92.16	92.30	97.63	97.57	92.13	97.16	94.91
+StarPO	**96.07**	95.65	96.55	95.36	95.85	95.85	97.88	97.58	93.80	96.45	96.11
OTHER SYSTEMS											
Tower+	93.02	94.89	93.13	95.50	94.93	93.78	96.46	96.04	**96.37**	95.43	94.96
GPT-4o	93.07	91.28	91.58	92.10	94.82	92.86	93.74	93.76	93.88	93.14	93.02
Deepseek-R1	93.56	**96.51**	**97.33**	**97.78**	96.26	94.12	90.39	93.25	93.08	97.09	94.93

Table 10: Performance in STAR score on News-Commentary test set, calculated by our proposed methods. **Bold** scores represent the global best performance and underlined scores represent the global second-best performance.

System	Ideal	Structural Deviations		
	1-to-1	1-to-0	0-to-1	Other
LLAMA-3.1-8B-INSTRUCT				
Base	92.59	2.08	0.17	4.15
+SFT	93.73	2.72	0.42	3.13
+CPO	95.42	0.33	0.75	3.50
+StarPO	95.79	1.95	0.14	2.12
QWEN-2.5-7B-INSTRUCT				
Base	95.35	1.91	1.31	1.43
+SFT	96.63	0.98	1.60	0.79
+CPO	97.46	0.35	1.52	0.67
+StarPO	98.43	0.68	0.00	0.89
QWEN3-4B-INSTRUCT				
Base	94.72	0.72	0.33	4.23
+SFT	95.36	0.80	0.00	3.84
+CPO	98.02	0.22	0.00	1.76
+StarPO	98.09	0.64	0.00	1.27
OTHER SYSTEMS				
Tower+	94.48	2.68	0.74	2.10
GPT-4o	92.91	2.25	2.89	1.95
Deepseek-R1	95.03	4.85	0.03	0.09

Table 11: Calculating STAR scores in Chinese ⇒ English language direction on News-Commentary test set by Gemini-2.5-Flash using prompts in Figure 4.

Metric	Spearman (ρ)	Pearson (r)	Kendall (τ)
Ground Truth	1.0000	1.0000	1.0000
LLM-based Evaluation			
LLM-judge (Appendix C)	0.7766	0.5947	0.6789
Existing Alignment Methods			
Align-then-Slide	0.4296	0.3928	0.2941
SEGALE	0.3912	0.4383	0.2940
Simple Heuristics			
Tokens Ratio	0.0024	0.0391	-0.0013
Sentence Ratio	0.2040	0.2281	0.1523
Sentence Count Difference	0.1528	0.1048	0.1089
Ours (STAR Variants)			
STAR (Default)	**0.7449**	**0.7284**	**0.6814**
w/o SaT (use Spacy)	0.5635	0.5777	0.4323
w/o LaBSE (use M3)	0.5864	0.6286	0.5281

Table 12: **Correlation analysis under the "Relaxed" setting.** In this variant, complex alignments (n-to-m) are counted as valid matches rather than penalties; scores decrease only for unaligned segments (insertions/deletions). This protocol is uniformly applied to all alignment baselines to ensure comparable conditions.

insertions (0-to-1, simulating hallucinations) and deletions (1-to-0, simulating omissions).

As shown in Table 13, under this regime, STAR significantly outperforms both existing alignment methods (Guo et al., 2025c; Wang et al., 2025e) and length-based heuristics. While simple heuristics struggle to pinpoint local errors and other alignment tools show limited correlation (e.g., $\rho =$ 0.28 for Guo et al. (2025c)), STAR achieves a strong correlation ($\rho = 0.67$), demonstrating its superior capability in detecting fundamental hallucinations and omissions even in the absence of complex structural permutations.

Metric	Spearman (ρ)	Pearson (r)	Kendall (τ)
Ground Truth	1.0000	1.0000	1.0000
LLM-based Evaluation			
LLM-judge (Appendix C)	0.8015	0.6675	0.6756
Existing Alignment Methods			
Align-then-Slide	0.2857	0.3248	0.2220
SEGALE	0.4165	0.4002	0.3591
Simple Heuristics			
Tokens Ratio	0.0071	0.0406	0.0022
Sentence Ratio	0.2558	0.3109	0.1957
Sentence Count Difference	0.2351	0.1637	0.1758
Ours (STAR Variants)			
STAR (Default)	**0.6686**	**0.6048**	**0.6328**
w/o SaT (use Spacy)	0.4382	0.5663	0.5020
w/o LaBSE (use M3)	0.4983	0.4672	0.4723

Table 13: **Correlation analysis on the Insertion/Deletion Only dataset.** This constructed dataset strictly excludes complex structural shifts (e.g., n-to-m alignments, swaps). It isolates the metrics' ability to detect pure insertions (0-to-1) and deletions (1-to-0).

System	Zh ⇔ En		Zh ⇔ De		Zh ⇔ Ru	
	⇒	⇐	⇒	⇐	⇒	⇐
LLAMA-3.1-INSTRUCT						
Base	30.12	57.60	16.40	28.65	18.37	31.70
+ SFT	70.94	61.46	42.68	60.62	78.63	50.25
+ CPO	72.80	74.15	73.41	80.96	75.81	87.34
+ StarPO	74.29	76.84	75.97	84.39	80.34	87.94
QWEN2.5-7B-INSTRUCT						
Base	60.30	75.01	69.90	42.02	64.10	82.58
+SFT	58.42	90.47	75.36	61.71	63.39	80.81
+CPO	71.16	89.04	72.75	75.09	81.81	81.39
+StarPO	74.64	91.48	**78.07**	86.23	**82.21**	87.47
QWEN3-4B-INSTRUCT						
Base	54.98	76.23	28.15	84.42	69.53	55.95
+SFT	77.63	89.55	43.49	81.99	71.58	81.56
+CPO	78.29	**90.11**	67.10	87.16	72.77	76.54
+StarPO	**83.74**	89.65	77.04	87.47	76.87	83.18
OTHER SYSTEMS						
Tower+	62.68	84.39	72.87	**86.46**	69.10	**91.44**
GPT-4o	65.09	82.74	70.31	72.73	77.58	66.88
Deepseek	64.00	66.57	69.84	64.52	67.78	62.72

Table 14: STAR scores on Guofeng test set.

H Comparison with Offline Preference Algorithms

To contextualize our method within the broader RLHF landscape, we compare STAR-Masked Preference Optimization against established offline algorithms, including DPO (Rafailov et al., 2023), SimPO (Meng et al., 2024), KTO (Ethayarajh et al., 2024), and ORPO (Hong et al., 2024). Specifically, we evaluate on the Chinese-to-English subset of the News-Commentary dataset. The results are presented in Table 16, indicating that vanilla DPO suffers from severe output collapse; however, adding an SFT loss allows DPO (+SFT) to match the performance of CPO. SimPO also exhibits occasional output collapse on specific entries. While KTO and SimPO prove to be effective, they yield slightly inferior results compared to CPO. Overall, our method demonstrates superior robustness and performance stability.

System	Zh ⇔ En		Zh ⇔ De		Zh ⇔ Ru	
	⇒	⇐	⇒	⇐	⇒	⇐
LLAMA-3.1-INSTRUCT						
Base	62.05	61.81	11.68	23.16	21.24	23.15
+ SFT	62.79	63.12	51.29	60.17	54.15	56.68
+ CPO	65.10	63.20	52.99	61.10	54.20	56.40
+ StarPO	65.85	62.47	64.17	55.20	54.74	56.94
QWEN2.5-7B-INSTRUCT						
Base	69.75	73.11	51.64	64.30	55.53	60.90
+SFT	67.58	74.97	51.28	63.88	55.63	61.19
+CPO	69.62	77.54	53.06	65.35	55.99	61.50
+StarPO	69.82	77.73	53.80	**66.69**	**56.33**	61.73
QWEN3-4B-INSTRUCT						
Base	68.04	78.67	53.51	61.49	52.40	57.63
+SFT	67.86	78.84	53.52	61.83	53.73	58.12
+CPO	70.50	78.95	54.10	62.33	54.06	63.07
+StarPO	**72.39**	**79.18**	54.46	62.94	54.29	63.72
OTHER SYSTEMS						
Tower+	64.78	64.61	**56.05**	65.41	52.49	**64.51**
GPT-4o	65.35	77.43	48.09	55.75	46.09	60.79
Deepseek	61.92	54.93	48.23	54.84	48.08	49.29

Table 15: COMETKiwi scores on Guofeng test set.

Method	COMET	COMETKiwi
DPO	36.86	22.61
DPO (*w.* SFT loss)	80.69	75.60
SimPO	70.48	70.94
ORPO	80.23	74.94
KTO	80.30	75.04
CPO (Standard)	81.10	75.24
Ours (StarPO)	**81.55**	**77.15**

Table 16: Comparison of different offline preference optimization algorithms.

I STAR Score Distribution

To further investigate the impact of structural constraints on preference pair construction, we visualize the score distributions of the standard STAR (Original) and its relaxed variant (STAR Relax) in Figure 6.

As illustrated in the top panel, a vast majority of samples are clustered at the perfect score of 1.0 in the STAR (Relax) settings. The bottom panel of Figure 6 provides a more granular view by excluding perfect 1.0 scores. Here, the contrast becomes more evident.

J Theoretical Justification and In-Depth Analysis of Masking Strategies

Below, we provide a theoretical justification for the experimental observations above. We analyze how the masking mechanism affects the reward

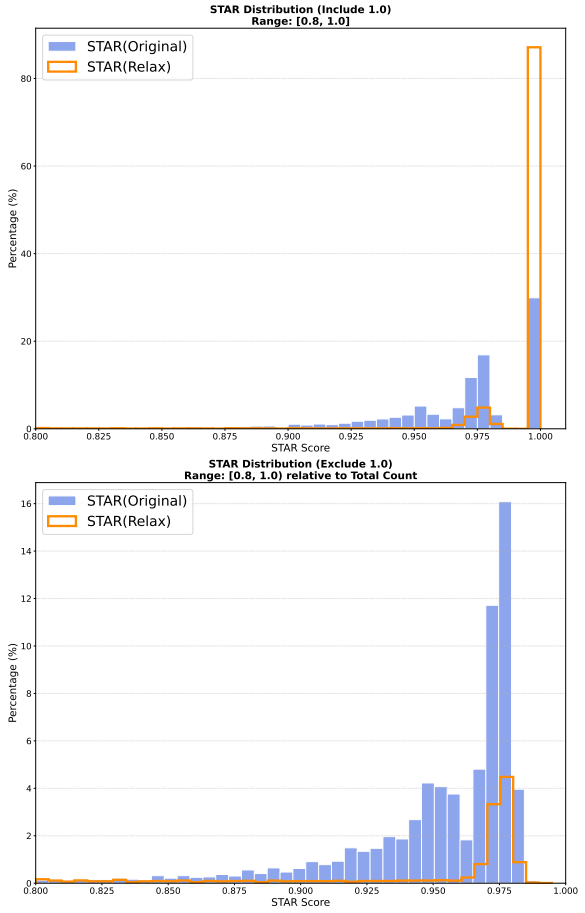

Figure 6: Histograms of metric scores for STAR (Original) and STAR (Relax). The **top** plot displays the full distribution including perfect matches (score=100 %) The **bottom** plot zooms in by excluding perfect matches, highlighting that STAR (Original) maintains a dense distribution of high-quality candidates, whereas STAR (Relax) has sparse coverage in the near-perfect region.

margins, the loss magnitude, and the gradient flow, specifically how it prevents the vanishing gradient problem often encountered in preference optimization.

1. Margin Scaling via Masking Let the standard log-likelihood margin between the preferred target y_w and the dis-preferred target y_l be denoted as Δ_{full}:

$$\Delta_{\text{full}}(x, y_w, y_l) = \log \pi_\theta(y_w|x) - \log \pi_\theta(y_l|x). \tag{7}$$

In the STAR-Masked objective, the likelihood is computed over a subset of sentences where the mask $\mathcal{M}(t_j) = 1$. Let $\mathcal{S}_{\text{mask}} \subset \{1, \ldots, n\}$ be the set of indices for sentences retained by the mask. The masked margin Δ_{STAR} is:

$$\Delta_{\text{STAR}}(x, y_w, y_l) = \log \pi_{\text{STAR}}(y_w|x) - \log \pi_{\text{STAR}}(y_l|x). \tag{8}$$

Since $\log \pi_{\text{STAR}}$ aggregates log-probabilities over a subset of tokens relative to the full document, the masked margin can be viewed as a scaled version of the full margin. Assuming the preference signal is distributed across the document, removing a portion of tokens (via $\mathcal{M}$) reduces the accumulated difference between y_w and y_l. Specifically, if the mask retains a ratio $\rho \in (0, 1)$ of the effective information:

$$|\Delta_{\text{STAR}}| \approx \rho \cdot |\Delta_{\text{full}}| < |\Delta_{\text{full}}|. \tag{9}$$

This derivation aligns with the empirical observation that reward margins decrease after introducing the mask, as shown in Figure 7.

2. Impact on Initial Loss Magnitude The preference loss component in CPO is defined as:

$$\mathcal{L}(\Delta) = -\log \sigma(\beta \cdot \Delta), \tag{10}$$

where Δ is the margin. The function $f(z) = -\log \sigma(z)$ is strictly monotonically decreasing. Assuming the model has a basic capability to distinguish y_w from y_l (i.e., $\Delta > 0$), the reduced margin caused by masking implies:

$$0 < \beta\Delta_{\text{STAR}} < \beta\Delta_{\text{full}}. \tag{11}$$

Due to the monotonicity of the loss function:

$$-\log \sigma(\beta\Delta_{\text{STAR}}) > -\log \sigma(\beta\Delta_{\text{full}}). \tag{12}$$

Thus, $\mathcal{L}_{\text{STAR-CPO}} > \mathcal{L}_{\text{CPO}}$ at the early stages of training. This theoretically confirms why the initial loss is higher and decays more slowly: the model perceives the "distance" between candidates as smaller, interpreting the optimization task as more difficult.

3. Gradient Saturation and Sustained Learning The efficacy of the optimization depends on the magnitude of the gradients. The gradient of the loss with respect to the model parameters θ is:

$$\nabla_\theta \mathcal{L} = \frac{\partial \mathcal{L}}{\partial \Delta} \cdot \nabla_\theta \Delta. \tag{13}$$

We focus on the scalar coefficient $\frac{\partial \mathcal{L}}{\partial \Delta}$, which modulates the strength of the update. For the CPO loss:

$$\frac{\partial \mathcal{L}}{\partial \Delta} = \frac{\partial}{\partial \Delta}\big(-\log \sigma(\beta\Delta)\big) = -\beta \cdot \big(1 - \sigma(\beta\Delta)\big). \tag{14}$$

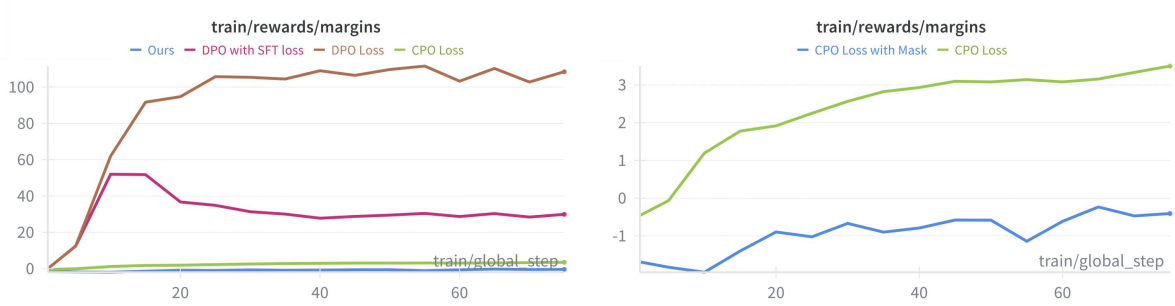

Figure 7: Analysis of training margins. The **left** panel compares the margins of Ours, DPO with SFT loss, standard DPO loss, and standard CPO loss. The **right** panel provides a zoomed-in view focusing on the comparison between standard CPO loss and Ours (indicated as "CPO Loss with Mask" in the legend).

We compare the gradient coefficients in two scenarios:

Scenario A: Full Objective (Standard CPO). If the model easily distinguishes y_w from y_l using simple patterns (e.g., trivial lexical differences in unmasked regions), Δ_{full} becomes large. As $\beta\Delta_{\text{full}} \to \infty$, $\sigma(\beta\Delta_{\text{full}}) \to 1$. Consequently, the gradient coefficient approaches zero:

$$|\nabla\mathcal{L}_{\text{CPO}}| \propto |1 - \sigma(\beta\Delta_{\text{full}})| \approx 0. \tag{15}$$

This leads to gradient saturation, where the model stops learning effectively even if structural errors persist.

Scenario B: Masked Objective (STAR-CPO). By masking out easy-to-align sentences (or random segments), we force Δ_{STAR} to be smaller. The value of $\sigma(\beta\Delta_{\text{STAR}})$ stays further from 1 (closer to the linear regime of the sigmoid function).

$$|1 - \sigma(\beta\Delta_{\text{STAR}})| > |1 - \sigma(\beta\Delta_{\text{full}})|. \tag{16}$$

Therefore, the masking mechanism acts as a regularizer that prevents the model from achieving a trivial margin on the training data. By artificially reducing the margin (Δ_{STAR}), the objective ensures that the gradient magnitude remains significant throughout the training process. This explains why, despite a slower decrease in loss, the model performs better in later steps: it avoids early saturation and continues to optimize the policy on the complex, structurally critical segments represented by the unmasked tokens.