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 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 to chunk-to-chunk (Chunk2Chunk) and 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
Doc2Doc translation in a single pass				
LLaMA-3.1-8B	92.59	2.08	1.17	4.16
Qwen-2.5-7B	95.35	1.91	1.31	1.43
Qwen-3-4B	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
Chunk2Chunk prompting LLaMA-3.1 *w. sentence boundary*				
MixSFT (Li et al., 2026a)	96.78	0.57	2.08	0.57
KFMT (Liu et al., 2025)	95.48	1.89	2.34	0.29
Ours (on Qwen2.5-7B)	98.43	0.68	0.00	0.89

Table 1: Alignment distribution (in %) for Zh ⇒ En on News-Commentary, computed by Gemini-2.5-Flash.

Most existing DocMT approaches implicitly avoid structural misalignment by adopting Sent2Sent or Chunk2Chunk paradigms, where translation is still performed at the individual sentence level given surrounding context (Wu et al., 2024a; Hu et al., 2025; Li et al., 2026b) or sliding chunks (Liu et al., 2025; Li et al., 2026a). While these localized paradigms successfully enforce sentence or block correspondence, they inherently limit global text planning and incur substantial computational costs, as source segments 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), 0-to-1 alignments (hallucinations) or even generation collapse (e.g., repetitive loops). Are non-1-to-1 alignments, where N source sentences correspond to M ($M \neq N$) target sentences, inherently unacceptable? Intu-

[1] See Appendix A for additional discussions.

itively, cross-lingual syntax often justifies merging, splitting, or reordering for fluency, ensuring the output feels more native and adheres to the idiomatic conventions of the target language. Therefore, a robust Doc2Doc system must therefore distinguish such legitimate restructuring from omissions or hallucinations.[2] Although splitting and merging are not categorized as pathological errors, effectively rectifying these structural nuances can yield substantial further performance gains.

As shown in our 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., 2026a). Notably, even strong proprietary models such as GPT-4o (OpenAI, 2024) exhibit non-trivial rates of structural mismatch in Doc2Doc and Chunk2Chunk settings (Liu et al., 2024; Shao et al., 2024a). These errors are particularly pronounced in long or complex documents. As input context chunk length scales up, the sentence alignment rate drops gradually (see Appendix C for a detailed quantitative analysis).

This vulnerability is 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. Although previous works (Hu et al., 2025; Domhan and Zhu, 2025) attempt to investigate these length-related translation failures, their analyses are largely confined to superficial comparisons of input and output lengths. Such macroscopic perspectives fail to capture the fine-grained structural breakdowns. Importantly, structural misalignment is largely invisible to standard training objectives and evaluation metrics (e.g. COMET (Rei et al., 2022)), which primarily focus on semantic adequacy and fluency. As a result, Doc2Doc systems may achieve high metric scores while still omitting or hallucinating.

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 $|\mathbf{s}_k|, |\mathbf{t}_k| \geq 1$).

[2] See Appendix B for case studies on distinguishing merging/splitting from omission/hallucination, and differentiating legitimate restructuring from those masking semantic loss.

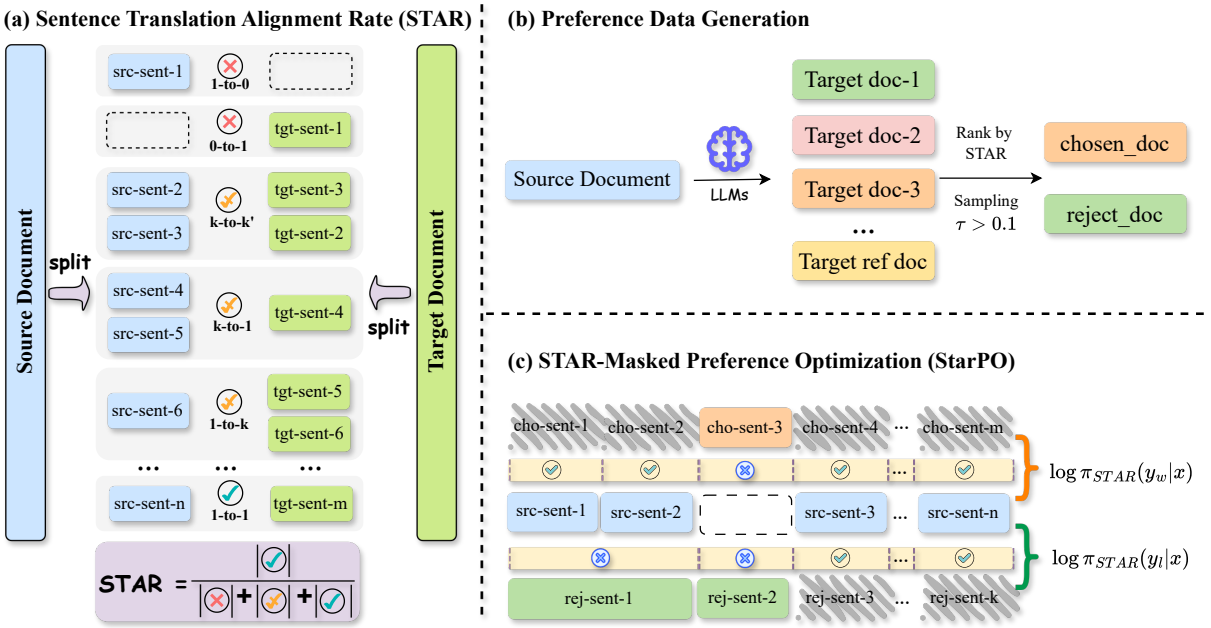

Figure 1: Overview of the proposed framework. (a) Sentence Translation Alignment Rate (STAR): Source and target documents are segmented and aligned to compute STAR, defined as the ratio of strict 1-to-1 alignments. (b) Preference Data Generation: Translation candidates are sampled and ranked by their 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.

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

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

We also introduce a variant, $\text{STAR}_{\text{relax}}(S,T)$, which treats $\mathcal{U}_{\text{complex}}$ as positive to exclusively penalize omissions and hallucinations while remaining tolerant of linguistically justified merging or splitting since complex alignments is also acceptable in real-world scenario:

$$\text{STAR}_{\text{relax}}(S,T) = \frac{|\mathcal{U}_{1:1}| + |\mathcal{U}_{\text{complex}}|}{|\mathcal{U}_{1:1}| + |\mathcal{U}_{1:0}| + |\mathcal{U}_{0:1}| + |\mathcal{U}_{\text{complex}}|}. \quad (2)$$

Thus, higher STAR indicates better sentence-level structural fidelity while lower STAR reflects deletions, insertions. Notably, STAR can also be computed directly via an LLM-as-a-judge approach, based on the four steps outlined above. See Appendix D 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 use GPT-4o (OpenAI, 2024) to generate 5 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 $\tau = 0.1$ [3].

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 optimiza-

[3]For more discussion of τ, see Appendix E.

tion (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:

$$\begin{aligned}\mathcal{L}_{\text{CPO}}(\pi_\theta) = &-\mathbb{E}_{\mathcal{D}}\Big[\log\sigma\big(\beta\log\pi_\theta(T_w|S) \\ &-\beta\log\pi_\theta(T_l|S)\big)\Big] \\ &-\mathbb{E}_{\mathcal{D}}\big[\log\pi_\theta(T_w|S)\big],\end{aligned} \tag{3}$$

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:

$$\mathcal{M}(t_j) = 1 - \mathbb{I}_{1:1}(t_j), \tag{4}$$

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:

$$\begin{aligned}\log\pi_{\text{STAR}}(T|S) = \sum_{j=1}^{n}\Bigg[&\mathcal{M}(t_j)\cdot \\ &\sum_{k=1}^{|t_j|}\log\pi_\theta\left(t_{j,k}|t_{<j}, t_{j,<k}, S\right)\Bigg],\end{aligned} \tag{5}$$

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:

$$\begin{aligned}\mathcal{L}_{\text{StarPO}}(\pi_\theta) = &-\mathbb{E}_{\mathcal{D}}\Big[\log\sigma\big(\beta\log\pi_{\text{STAR}}(T_w|S) \\ &-\beta\log\pi_{\text{STAR}}(T_l|S)\big)\Big] \\ &-\mathbb{E}_{\mathcal{D}}\big[\log\pi_{\text{STAR}}(T_w|S)\big].\end{aligned} \tag{6}$$

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 web novels and news from WMT. Specifically, we use News-Commentary v18.1 from WMT25 for English (En) ⇔ { Chinese (Zh), German (De), Russion (Ru), Spanish (Es)} and Chinese (Zh) ⇔ German (De) and Guofeng (Wang et al., 2024)[4] from WMT25[5] for Chinese (Zh) ⇔ {English (En), German (De), Russian (Ru)} translation. We use the training sets to fine-tune the LLMs and to construct preference pairs. Detailed dataset statistics are provided in Appendix F.

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

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**) following Li et al. (2026a), we use our full training set, 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).

[4] github.com/longyuewangdcu/GuoFeng-Webnovel/
[5] www2.statmt.org/wmt25/translation-task.html
[6] huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[7] huggingface.co/Qwen/Qwen2.5-7B-Instruct
[8] 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**
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		De ⇔ En		De ⇔ Zh		Ru ⇔ En		En ⇔ Es		Avg.
	⇒	⇐	⇒	⇐	⇒	⇐	⇒	⇐	⇒	⇐	
LLAMA-3.1-8B-INSTRUCT											
Base	19.25	20.03	36.37	27.08	22.89	11.56	40.31	22.21	42.15	40.25	28.21
+ SFT	21.55	39.36	44.20	28.54	30.62	15.43	41.65	23.78	42.60	43.49	33.12
+ CPO	30.98	38.77	44.02	28.27	34.98	16.69	42.11	24.45	43.14	43.52	34.69
+ StarPO	32.88	40.20	**44.36**	28.77	35.90	**17.94**	**42.45**	25.72	43.99	43.68	**35.59**
QWEN2.5-7B-INSTRUCT											
Base	30.14	41.63	41.19	24.28	36.29	13.29	40.65	21.25	41.28	43.03	33.30
+ SFT	29.62	42.27	42.97	25.34	35.08	14.45	41.63	24.05	41.57	43.59	34.06
+ CPO	30.90	43.34	43.71	25.78	**37.46**	14.52	41.87	24.10	41.71	43.13	34.65
+ StarPO	**35.23**	**44.12**	44.34	26.52	36.34	14.92	42.05	25.04	42.14	**44.02**	35.47
OTHER SYSTEMS											
Tower+	33.37	41.89	43.16	28.52	34.22	15.08	41.97	26.95	42.85	43.16	35.12
GPT-4o	27.70	42.06	42.83	29.62	35.38	12.70	41.28	**28.55**	**44.24**	43.12	34.75
Deepseek-R1	32.33	39.27	41.68	**29.83**	33.59	14.00	40.44	27.28	44.06	41.64	34.41

Table 3: Performance in d-BLEU score on the News-Commentary test set. Unlike dCOMET, d-BLEU evaluates the document as a single continuous string without relying on sentence alignment tools. 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 ≥ 3.35 for BLEU.

Metrics. We evaluate translation quality using document-level COMET (dCOMET) (Vernikos et al., 2022), computed by `wmt22-comet-da` (Rei et al., 2022). 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), using our settings in $\text{STAR}_{\text{relax}}$, we assign a score of 0 COMET to unaligned segments, including omissions (1-to-0) and hallucinations (0-to-1), while complex mappings (1-to-k, k-to-1, k-to-k') **retain** their computed COMET scores. To provide a more holistic evaluation, we use d-BLEU, which bypasses sentence-splitting tools by treating the entire document as a single continuous string for both input and output. This ensures that the evaluation remains uninfluenced by potential sentence-boundary artifacts. See Appendix I for our proposed STAR and $\text{STAR}_{\text{relax}}$ scores.

3.2 Experimental Results

Table 2 and Table 3 present the results on the News-Commentary dataset.[9] Across all backbones , supervised fine-tuning (+SFT) yields consistent and substantial improvement over the base models on most language pairs. Building on this, stan-

[9]See Appendix J for results on the Guofeng dataset.

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.* BLEU)	77.81	79.52
GRPO (*w.* COMET + STAR)	77.76	79.76
GSPO (*w.* COMET)	77.94	79.77
GSPO (*w.* STAR)	80.16	80.00
GSPO (*w.* BLEU)	77.96	79.81
GSPO (*w.* COMET + STAR)	77.72	79.97
Our Variants & Ablation Studies		
SFT on preference data	80.41	80.02
STAR (Relaxed)	80.09	79.59
Random Mask (Sentence level)	81.18	80.05
Random Mask (Token level)	81.17	80.06
CPO	81.10	79.94
Ours (StarPO)	**81.55**	**80.11**

Table 4: COMET scores comparison against alternative optimization strategies. Word-level Coverage is calculated via WSPAlign (Wu et al., 2024c, 2023).

dard CPO (+CPO) further enhances translation quality. Our proposed StarPO consistently outperforms standard CPO 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 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.

3.3 Comparison with Alternative Optimization Baselines

To further illustrate the effectiveness of StarPO, we compare it against a diverse set of alternative optimization baselines. We use LLaMA-3.1-8B-Instruct on Zh ⇔ En as a representative setting and report the results in Table 4.

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 4, 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, STAR and BLEU scores into the reward function for these baselines. As observed in the middle section of Table 4, 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 analyses and comparisons with other standard offline preference optimization algorithms (e.g., DPO (Rafailov et al., 2023), SimPO (Meng et al., 2024)) are provided in Appendix K.

Ablation Studies. To further dissect the source of our gains, we examine several specific variants in the bottom section of Table 4: **(1) SFT vs. Preference Optimization:** Training solely on preferred responses (y_w) yields strong performance, yet our full framework consistently outperforms this SFT baseline , proving that the contrastive signal in preference optimization is more effective than simple supervised imitation. **(2) Strict vs. Relax STAR:** Using the relax STAR metric degrades performance, as it dilutes the discriminative power needed to establish a sufficient score margin for data selection during training (see Appendix L for score distribution analysis). **(3) Alignment-aware vs. Random Masking:** We further investigate the effect of loss masking by randomly masking out 90% of tokens[10] at both the sentence and token levels. While random token- or sentence-level masking acts as a helpful regularizer (Gu et al., 2025), StarPO's semantic-aware masking consistently achieves superior results. This demonstrates that focusing the optimization objective on structurally misaligned segments provides a more effective learning signal than random selection.[11]

[10] See Appendix M for the distribution of unmasked token ratios during training.

[11] See Appendix N for theoretical analysis.

Methods	Correlation ($\rho, \uparrow$)	
	Strict 1-to-1	**Relaxed N-to-M**
Existing Alignment Methods		
Align-then-Slide	0.3804	0.5168
SEGALE	0.4193	0.4760
Simple Metrics		
Token Count Ratio	0.0218	0.0211
Sentence Count Ratio	0.1218	0.1780
Sentence Count Difference	0.0649	0.1647
Ours (STAR Variants)		
STAR (Section 2.1)	**0.5808**	**0.5774**
w/o SaT (use Spacy)	0.5644	0.5663
w/o LaBSE (use M3)	0.5104	0.5479

Table 5: Spearman correlation (ρ) between LLM-annotated alignment quality and various evaluation methods (including STAR variants, Align-then-Slide (Guo et al., 2025c), SEGALE (Wang et al., 2025e), and length-based metrics) under both strict 1-to-1 and relaxed N-to-M matching criteria.

4 Discussions

4.1 Sensitivity to Structural Pathologies

To verify STAR's capability in identifying structural pathologies, we evaluate its performance. Starting from WMT24 (Kocmi et al., 2024a) test set,[12] we introduce several LLMs including GPT-4o, Gemini-1.5-Flash, Hunyuan(Sun et al., 2024), Qwen3-235B[13], and DeepSeek-R1 to produce translations across several language pairs: En $\Leftrightarrow$ {De, Zh, Ru, Es}. To establish a reliable ground truth, we employ Gemini-2.5-Pro to annotate the structural alignment score (i.e. LLM implementation of STAR). We then compute Spearman (ρ) between the metric scores and the labels in two settings: strict mode and relax mode. Specifically, the strict mode treats only 1-to-1 mappings as valid alignments, enforcing a rigid sentence-level correspondence. In contrast, the relax mode accounts for "reasonable restructuring" by accepting linguistically justified N-to-M mappings (e.g., merging or splitting) as valid, thereby penalizing only the pathological 1-to-0 omissions and 0-to-1 hallucinations. The results are shown in Table 5.

Comparison with Existing Alignment Methods. Both Align-then-Slide (Guo et al., 2025c) and SEGALE (Wang et al., 2025e) utilize sentence alignment between source and target text for document-level evaluation. We compute STAR scores using their intermediate alignment results for a fair comparison. As shown in Table 5, our implementation achieves superior correlation with structural noise. This highlights that precise, fine-grained sentence alignment is a critical prerequisite for reliable document-level structural assessment.

Comparison with length-based Metrics. We compare STAR with length-based metrics in prior works for detecting omissions, hallucinations or collapse (Peng et al., 2025a; Guerreiro et al., 2023; Shao et al., 2024a; Hu et al., 2025; Domhan and Zhu, 2025), including *Token Count Ratio*, *Sentence Count Ratio*, and *Sentence Count Difference*. As shown in Table 5, these metrics exhibit weak correlations ($\rho < 0.2$), failing to detect structural noise when overall document length is preserved.

Ablation Study on STAR Components. We verify the robustness of STAR by swapping its components. Replacing SaT with Spacy[14] or switching the LaBSE encoder to M3 (Chen et al., 2024) results only marginal performance fluctuations. These results demonstrate that STAR's effectiveness is relatively robust.

4.2 Analysis of Structural Alignment and Semantic Accuracy

While complex alignments (k-to-k', including merging and splitting) are sometimes feasible, they frequently mask local semantic distortions or partial omissions. To investigate this, we conduct a fine-grained COMET analysis on Zh $\Rightarrow$ En, categorizing source sentences into three groups: **Pathological Errors** (1-to-0/0-to-1), **Complex Restructuring** (merging/splitting/swapping), and **Preserved** 1-to-1 mappings.

Table 6 shows that StarPO's improvement stems from mitigating pathological errors and, more significantly, rectifying *Complex* restructuring. While SFT offers marginal gains, StarPO substantially boosts the COMET score for complex segments (71.87 $\rightarrow$ 81.52), confirming that enforcing structural 1-to-1 correspondence recovers semantic details typically lost in merging or splitting.

Pattern	Pathological		Complex		1-to-1		Overall
	%	COMET	%	COMET	%	COMET	COMET
Base	3.25	0.00	4.16	71.87	92.59	75.78	72.23
SFT	3.14	0.00	3.13	74.81	93.73	79.45	76.81
StarPO	2.09	0.00	2.12	81.52	95.79	83.33	81.55

Table 6: Fine-grained analysis on Zh $\Rightarrow$ En.

[12] github.com/wmt-conference/wmt24-news-systems
[13] huggingface.co/Qwen/Qwen3-235B-A22B
[14] https://github.com/explosion/spaCy

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

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., 2025), we use Gemini-2.5-Flash for all systems. The results of News-Commentary Zh ⇒ En are shown in Table 7. 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 selection (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), self-refinement (Koneru et al., 2024; Li et al., 2025b) and employing memory-based agents (Wang et al., 2025f; Guo et al., 2025b). Fine-tuning relies on various data construction strategies (Li et al., 2026b; Lyu et al., 2024; Wu et al., 2024a; Zhang et al., 2023b; Stap et al., 2024), with the Tower series (Alves et al., 2024; Rei et al., 2024, 2025) being a representative case. Mąka et al. (2025b,a); Mohammed and Niculae (2025); Choudhary et al. (2025); Li et al. (2026b) further analyze the mechanics of context usage for context-aware translation.

Doc2Doc approaches aim for holistic translation through long-context training (Pang et al., 2025; Li et al., 2026a), iterative or agentic refinement (Dong et al., 2025; Li et al., 2025b; 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., 2026; 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, 2025). Such online frameworks is also compatible for training large reasoning models (Feng et al., 2025a,b; Wang et al., 2025b,c,d). 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 approaches are proposed to construct and leverage these datasets (Sun et al., 2025a; Cui et al., 2025; Tang 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 utilizing dynamic masking to target omissions and hallucinations. Experiments demonstrate StarPO enables compact models to surpass massive proprietary systems in translation quality while significantly improving token efficiency. This work establishes a robust paradigm for "in-one-go" 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.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of ACL*, pages 2318–2335.

Zhi-Yuan Chen, Hao Wang, Xinyu Zhang, Enrui Hu, and Yankai Lin. 2025. 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. Model alignment as prospect theoretic optimization. In *Proceedings of ICML*, pages 12634–12651.

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. *CoRR*, abs/2505.19714.

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 *Proceedings of ICLR*.

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.

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. In *Proceedings of NLPCC*, page 228–240.

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.

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.

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 ACL*, pages 1999–2014.

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.

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. 2026a. Enhancing document-level translation of large language model via translation mixed instructions. *Neurocomputing*, 664:132041.

Ying Li, Xinglin Lyu, Junhui Li, Jinlong Yang, Hengchao Shang, Min Zhang, Shimin Tao, and Daimeng Wei. 2026b. Cross-preference learning for sentence-level and context-aware machine translation. *CoRR*, abs/2603.25183.

Zongyao Li, Zhiqiang Rao, Hengchao Shang, Jiaxin Guo, Shaojun Li, Daimeng Wei, and Hao Yang. 2025b. 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 mt with multi-knowledge fusion. In *Proceedings of NLPCC*, page 175–187. Springer-Verlag.

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.

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.

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. 2026. Fine-grained reward optimization for machine translation using error severity mappings. *Transactions of the Association for Computational Linguistics*, 14:733–754.

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. 2022. 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.

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 EMNLP: System Demonstrations*, pages 256–263.

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.

Xingwu Sun, Yanfeng Chen, Yiqing Huang, Ruobing Xie, Jiaqi Zhu, Kai Zhang, Shuaipeng Li, Zhen Yang, Jonny Han, Xiaobo Shu, Jiahao Bu, Zhongzhi Chen, Xuemeng Huang, Fengzong Lian, Saiyong Yang, Jianfeng Yan, Yuyuan Zeng, Xiaoqin Ren, Chao Yu, and 87 others. 2024. Hunyuan-large: An open-source moe model with 52 billion activated parameters by tencent. *CoRR*, abs/2411.02265.

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. *CoRR*, abs/2504.10187.

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.

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 *Proceedings of 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.

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.

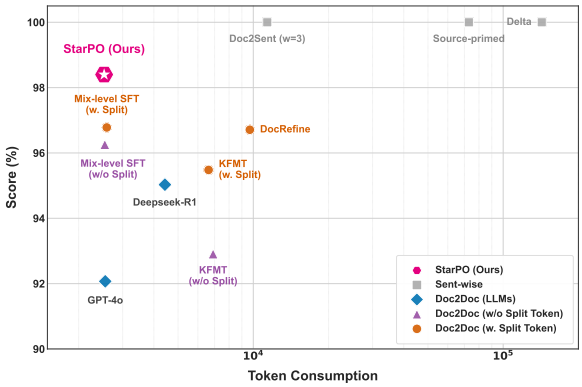

Figure 2: **Alignment score vs. Token Consumption.** We compare our proposed **StarPO** against various baselines, including sentence-level systems, document-level baselines, and proprietary LLMs (e.g., GPT-4o, Deepseek-R1). Note that the x-axis is plotted on a logarithmic scale (10^n).

A Efficiency Analyses

We evaluate our framework against several representative document translation paradigms. Figure 2 compares the alignment score vs. token consumption among several typical Doc2Sent and Doc2Doc systems. Doc2Sent (w=3) (Wu et al., 2024a; Lyu et al., 2024; Koneru et al., 2024; Cui et al., 2024) utilizes a standard sliding window of three preceding sentences for context, while Doc2Sent (Source-primed) (Hu et al., 2025) treats translation as a sequential multi-turn dialogue. Doc2Sent (DelTA) (Wang et al., 2025f) employs an agentic framework for autonomous context management. Among document-level models, Doc2Doc (KFMT) (Liu et al., 2025) and Doc2Doc (Mix-level SFT) (Li et al., 2026a) use multi-turn interactions or specialized fine-tuning, evaluated both with and without sentence delimiters. Doc2Doc (DocRefine) (Dong et al., 2025) focuses on iterative multi-stage refinement to improve coherence. For large-scale baselines, we include Doc2Doc (DeepSeek-R1) to represent reasoning-based models and Doc2Doc (GPT-4o) as a high-performance proprietary benchmark. Finally, Doc2Doc (StarPO) is our proposed framework, which uses STAR-curated preference pairs to optimize for both structural alignment and semantic fidelity.

B Detailed Case Study

While STAR prioritizes 1-to-1 alignment, Figure 3 illustrates that structural deviations range from pathological errors to valid linguistic adaptations. We categorize these patterns based on their semantic impact:

- **Pathological Deviations:** 1-to-0 (Omission) and 0-to-1 (Hallucination) instances that result in clear fidelity loss (red).
- **Structural Adaptations:** k-to-1 (Merging), 1-to-k (Splitting), and k-to-k' (Swapping) patterns that often optimize cohesion or readability (green).

As detailed in the figure's notes, even valid adaptations (Cases 3a, 4a, 5a) may harbor localized semantic inaccuracies (Cases 3b, 4b, 5b), necessitating a nuanced evaluation that distinguishes stylistic alignment from actual information loss.

C Impact of Input Context Length on Structural Alignment

To further investigate the relationship between document length and structural fidelity in Doc2Doc translation, we evaluated the Sentence Translation Alignment Rate (STAR) across varying input context lengths using LLaMA-3-8B-Instruct on News-Commentary test set on Zh $\Rightarrow$ En translation. As illustrated in Figure 4, there is a clear negative correlation between the input length (measured in tokens) and the alignment quality.

In the standard sentence-to-sentence (Sent2Sent) translation paradigm, the model naturally achieves a perfect 100% STAR score. However, as the input context expands to 256 tokens, the alignment rate sharply drops to 96.69%. This highlights the critical need for explicit structural alignment objectives, such as our proposed StarPO, particularly for long-document translation.

D 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 5. This prompt is designed to simulate human-level judgment on document structure.

E Discussion of Threshold τ

To investigate the impact of the margin threshold τ used in preference data construction, we conduct a sensitivity analysis across both Zh $\Rightarrow$ En and En $\Rightarrow$ Zh directions. As shown in Table 8, the performance variation remains minimal (less than 0.6 points) as τ ranges from 0.00 to 0.15. This stability suggests that StarPO is robust to the choice

Cases

1 1-to-0(Omission)
The proposal was accepted unanimously. However, concerns regarding the budget remain.
提案全票通过。

2 0-to-1 (Hallucination)
The weather forecast predicts rain tomorrow.
天气预报显示明天有雨。 建议居民待在室内。

3a k-to-1 (Merging) with semantic persist ☑
But will that still be true after six months? Or nine?
但六个月，或者九个月后还会是这样吗？

3b k-to-1 (Merging) with semantic loss ✗
Second, it seems likely that nominal growth in the US will rise, though the underlying ***mix*** of inflation and real growth is yet to be determined. This matters, because it will shape the US Federal Reserve's response, affecting asset prices in America and beyond.
Base :第二，似乎美国的名义经济增长率会提高(然而本文写作时尚未可知)，但实际增长率是否跟随增长率提升至关重要，因为这决定了美联储如何做出反应，从而影响美国及其以外地区的资产价格。
Ours: 其次，美国的名义增长有可能上升，尽管基础的通胀和实际增长的***混合***尚未确定。这很重要，因为它将影响美国联邦储备委员会的反应，影响美国和其他地区的资产价格。
Ref: 其次，看起来美国的名义增长很有可能会上升，尽管通胀和真实增长的决定性***组合***情况尚不清楚。这很重要，因为会影响到美联储的反应，进而影响美国和其他国家的资产价格。

Note: Merging ≠ Omission: Sentence count reduction via k-to-1 alignment can effectively enhance logical cohesion and conciseness in different languages (case 3a) , distinguishing it from severe content loss where full sentences are omitted, though local semantic loss may still persist within the restructured output (case 3b).

4a 1-to-k (Splitting) with semantic persist ☑
Diversity, which cannot be imposed on anyone, must be demanded by society and sustained through the establishment of necessary long-term institutions.
多元化不能强加给任何人。社会必须要求多元化，并建立起必要的长期机构来维持多元化。

4b 1-to-k (Splitting) with semantic loss ✗
This new pattern - which persisted through the 1908s and 1990s, and accelerated ***after 2000*** - caused inequality to rise sharply, weakening the foundations of globalization.
Base : 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。
Ours: 这一模式——它在20世纪80年代和90年代持续存在，***2000年后***加速了——导致不平等急剧上升，削弱了全球化的基础。
Ref: 这一新模式——它贯穿了20世纪八九十年代，***进入21世纪后***更是不断加速——导致了不平等性急剧升高，破坏了全球化的基础。

Note: Splitting ≠ Hallucination: Sentence count increase via 1-to-k alignment can effectively enhance readability (case 4a), distinguishing it from hallucinating full sentences, though local semantic loss may still persist (case 4b).

5a k-to-k' (Swapping) with semantic persist ☑
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。

5b k-to-k' (Swapping) with semantic loss ✗
Germany stands as a significant exception to this pattern. Even before joining the Eurozone, the country was accustomed to low borrowing costs.
Base: 在进入欧元区之前，德国就习惯于低借贷成本，因此债务沉重。德国是这一模式的重要例外。
Ours: 这一模式的重要例外是德国，它在加入欧元区之前就习惯于低借贷成本。
Ref: 这一模式的重要例外是德国，德国在加入欧元区前便已习惯了低借贷成本。

Note: Swapping ≠ Error: *Semantic elements are redistributed across sentence boundaries*. This requires treating the sequence as a single semantic block rather than isolated units (case 5a); otherwise, they may be wrongly penalized as local omission in one sentence and hallucination in another, though local semantic loss may still persist (case 5b).

Figure 3: Case study of structural alignment patterns. Green text indicates acceptable linguistic adaptations (e.g., merging for cohesion or splitting for clarity), while red text highlights genuine translation errors. These two scenarios must be carefully distinguished to avoid misjudging stylistic alignment as semantic loss.

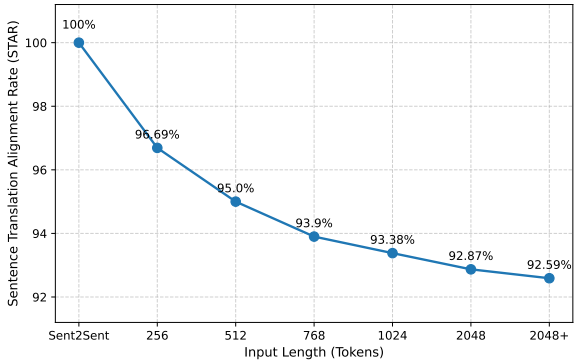

Figure 4: Sentence Translation Alignment Rate (STAR) across varying input context lengths from Sent2Sent to more than 2048 tokens. The results are evaluated on Zh ⇒ En translation of the News Commentary dataset using LLaMA-3.1-8B-Instruct.

τ	**Zh ⇒ En**	**En ⇒ Zh**
0.00	81.04	80.01
0.05	81.15	79.96
0.10	81.55	80.11
0.15	81.38	80.08

Table 8: Effect of τ on Zh ⇔ En translations.

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 9: Statistics of the News-Commentary dataset.

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 10: Statistics of the Guofeng dataset.

of τ and does not require extensive hyperparameter tuning. We select $\tau = 0.10$ as the default for all experiments, as it yields a marginal but consistent advantage in our pilot studies.

F 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 9 and Table 10, 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).

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.

G Implementation Details

We implement our experiments in ms-swift (Zhao et al., 2025)[15]. During fine-tuning, we adapt LoRA (Hu et al., 2021). We set LoRA rank to 8 and LoRA alpha to 16 , respectively. The models 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. Regarding evaluation efficiency, our proposed STAR metric is computationally efficient, achieving a throughput of approximately 3 samples per second. During inferencing, we set temperature to 0.3, beam size to 1 in vllm[16] framework. The specific prompt template used in our experiments is illustrated in Figure 6.

H Results on Qwen-3-4B-Instruct

Table 11 and Table 12 show the performance of Qwen3-4B-Instruct (Yang et al., 2025) on the News-Commentary test set.

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 5: 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.

I STAR Scores

The structural fidelity of these systems, as measured by our proposed STAR score, is detailed in Tables 13 and 14. Additionally, we report the alignment results on the News-Commentary Zh ⇒ En dataset evaluated by Gemini-2.5-Flash using the prompt from Appendix D, as shown in Table 15.

J Results on Guofeng

Table 16 underscores StarPO's robustness in literary translation. 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 1-to-1 constraints during training, our evaluation remains inclusive of valid complex mappings. StarPO's training constraints prevent pathologies like language mismatches without stifling this flexibility, as evidenced by restoring coherence where LLaMA-3.1-base fails on Zh⇒De (52.31 COMET). 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.

[15] https://github.com/modelscope/ms-swift
[16] https://github.com/vllm-project/vllm

System	Zh ⇔ En		De ⇔ En		De ⇔ Zh		Ru ⇔ En		En ⇔ Es		Avg.
	⇒	⇐	⇒	⇐	⇒	⇐	⇒	⇐	⇒	⇐	
QWEN-3-4B-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

Table 11: Performance in dCOMET score on the News-Commentary test set on Qwen3-4B-Instruct.

System	Zh ⇔ En		De ⇔ En		De ⇔ Zh		Ru ⇔ En		En ⇔ Es		Avg.
	⇒	⇐	⇒	⇐	⇒	⇐	⇒	⇐	⇒	⇐	
QWEN-3-4B-INSTRUCT											
Base	27.71	41.55	42.22	26.01	31.68	14.27	38.34	25.00	40.58	43.91	33.13
+ SFT	30.35	42.07	43.61	27.09	33.72	14.79	41.41	27.37	42.79	44.09	34.73
+ CPO	33.68	42.54	43.80	27.24	34.41	15.33	41.80	28.11	42.68	44.46	35.41
+ StarPO	35.93	43.42	43.41	28.11	33.66	15.61	42.09	28.12	42.62	45.14	35.81

Table 12: Performance in d-BLEU score on the News-Commentary test set on Qwen3-4B-Instruct.

Prompts for In-one-go Document-level Translation

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

```
<src_doc>
```

Figure 6: **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.

K Details in Preference Optimization

We first highlight the practical advantages of STAR over existing model-based metrics. While employing COMET as a reward function for GRPO/GSPO training is computationally expensive, it also exhibits training instability, with the reward loss frequently fluctuating around zero. Furthermore, using COMETKiwi as a reward signal may even lead to off-the-target language generation, where the model erroneously translates the source into an unintended language. In contrast, STAR offers a significantly more efficient and robust alternative, enabling rapid reward calculation (approx. 3 samples/s) while ensuring stable convergence during preference optimization.

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 17, 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.

L 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 7.

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 7 provides a more granular view by excluding perfect 1.0 scores. Here, the contrast becomes more evident.

System	Zh ⇔ En		De ⇔ En		De ⇔ Zh		Ru ⇔ En		En ⇔ Es		Avg.
	⇒	⇐	⇒	⇐	⇒	⇐	⇒	⇐	⇒	⇐	
LLAMA-3.1-8B-INSTRUCT											
Base	90.31 96.44	87.21 93.65	92.52 94.43	95.41 95.92	78.96 93.55	82.79 89.37	97.05 98.51	93.92 96.46	90.97 96.98	94.06 95.86	90.32 95.12
+ SFT	93.52 95.00	91.75 95.05	95.77 96.39	95.60 97.03	80.39 94.38	93.58 96.28	96.97 98.58	97.58 97.82	77.13 96.04	95.64 96.27	91.79 96.28
+ CPO	95.00 97.05	91.72 96.62	95.99 96.90	95.52 98.76	79.83 97.93	90.41 97.42	97.30 98.91	97.65 97.88	90.64 91.87	97.50 97.82	93.16 97.12
+StarPO	95.62 98.64	93.36 98.31	96.80 97.08	95.59 98.66	80.80 98.36	**97.12** 97.54	97.27 98.90	**98.01** 98.08	90.13 92.45	**97.63** 98.01	94.23 97.60
QWEN2.5-7B-INSTRUCT											
Base	94.97 96.42	93.92 95.72	95.85 96.72	95.10 96.11	61.80 72.68	92.03 94.38	98.05 98.46	86.87 92.16	96.02 96.77	97.33 97.76	91.19 93.72
+SFT	95.51 96.90	92.96 98.85	94.54 96.88	94.45 96.66	88.42 92.00	93.25 97.20	97.49 98.82	95.55 98.99	81.10 95.38	96.80 97.32	93.01 96.90
+CPO	95.40 98.93	91.09 **99.00**	95.89 98.96	95.36 96.63	95.92 96.04	92.49 97.12	97.64 99.08	95.54 99.32	92.83 92.94	97.56 99.15	94.97 97.72
+StarPO	**96.07** **99.29**	95.67 98.77	96.55 99.45	95.36 96.71	**96.83** **99.01**	93.12 97.35	**98.15** **99.68**	97.71 **99.74**	96.21 97.01	97.37 99.60	**96.30** 98.66
QWEN3-4B-INSTRUCT											
Base	94.67 98.73	93.41 94.02	96.33 98.08	96.45 96.81	82.66 93.10	91.01 94.79	97.69 98.08	96.92 97.12	96.12 96.21	91.38 94.95	93.66 96.19
+SFT	95.02 98.90	92.92 98.91	96.48 99.11	91.05 98.48	83.03 96.44	91.26 96.76	97.48 98.97	95.35 99.63	95.36 97.14	96.12 99.58	93.41 98.39
+CPO	95.01 98.93	93.24 98.90	96.60 **100.0**	95.33 **98.99**	92.16 96.74	92.30 97.57	97.63 99.05	97.57 97.91	92.13 96.58	97.16 **99.71**	94.91 98.44
+StarPO	**96.07** 99.13	95.65 98.99	96.55 **100.0**	95.36 98.62	95.85 98.45	95.85 **98.92**	97.88 99.35	97.58 99.70	93.80 **98.89**	96.45 99.62	96.10 **99.17**
OTHER SYSTEMS											
Tower+	93.02 94.15	94.89 95.45	93.13 95.79	95.50 96.50	94.93 96.03	93.78 96.02	96.46 98.04	96.04 98.04	**96.37** 98.16	95.43 97.78	94.96 96.60
GPT-4o	93.07 94.66	91.28 96.76	91.58 97.38	92.10 97.26	94.82 98.23	92.86 96.39	93.74 97.53	93.76 97.46	93.88 97.54	93.14 96.80	93.02 97.00
Deepseek-R1	93.56 94.04	**96.51** 97.81	**97.33** 97.44	**97.38** 97.88	96.22 96.26	94.12 98.03	90.39 92.49	93.25 97.05	93.08 97.22	97.09 97.12	94.89 96.53

Table 13: Performance in STAR scores on News-Commentary test set. In each cell, the top value is the **Strict** score and the bottom value (in gray) is the **Relaxed** score.

M Masked Token Distribution

Figure 8 illustrates the probability density distribution of the unmasked token ratio.

N 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 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:

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

Since $\log \pi_{\text{STAR}}$ aggregates log-probabilities

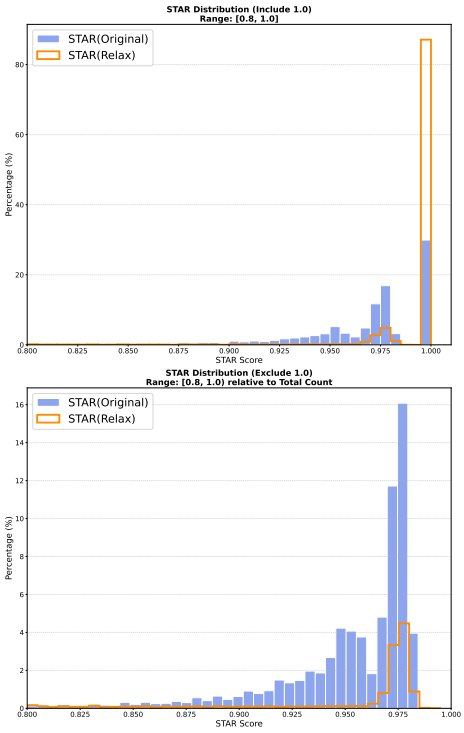

Figure 7: Histograms of metric scores for STAR (Original) and STAR (Relax). The **top** plot displays the full distribution including perfect matches (score=100 %) The **bottom** plot zooms in by excluding perfect matches, highlighting that STAR (Original) maintains a dense distribution of high-quality candidates, whereas STAR (Relax) has sparse coverage in the near-perfect region.

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 9.

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

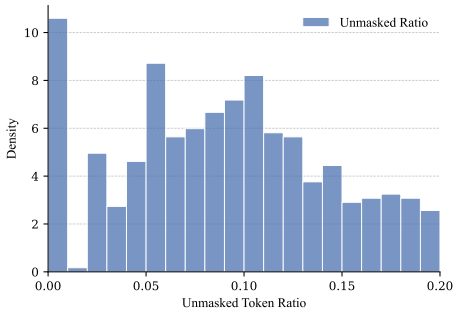

Figure 8: Frequency distribution of unmasked tokens. To provide a clear view of the token preservation statistics, the distribution is normalized

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:

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

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.

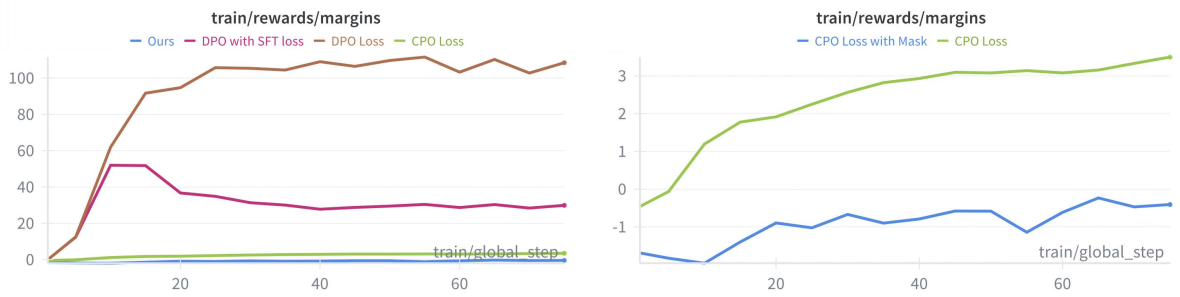

Figure 9: Analysis of training margins. The **left** panel compares the margins of Ours, DPO with SFT loss, standard DPO loss, and standard CPO loss. The **right** panel provides a zoomed-in view focusing on the comparison between standard CPO loss and Ours (indicated as "CPO Loss with Mask" in the legend).

Scenario B: Masked Objective (StarPO). 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).

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

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.

System	Zh ⇔ En		Zh ⇔ De		Zh ⇔ Ru	
	⇒	⇐	⇒	⇐	⇒	⇐
LLaMA-3.1-Instruct						
Base	30.12 38.84	57.60 72.33	16.40 23.30	28.65 35.57	18.37 35.53	31.70 44.35
+ SFT	70.94 80.51	61.46 83.67	42.68 78.67	60.62 69.59	78.63 80.89	50.25 62.60
+ CPO	72.80 80.74	74.15 77.01	73.41 80.06	80.96 85.98	75.81 80.03	87.34 98.58
+ StarPO	74.29 82.37	76.84 83.05	75.97 79.39	84.39 88.32	80.34 85.40	87.94 97.39
Qwen2.5-7B-Instruct						
Base	60.30 95.83	75.01 91.61	69.90 93.56	42.02 91.93	64.10 92.94	82.58 91.63
+SFT	58.42 94.71	90.47 93.79	75.36 95.44	61.71 88.59	63.39 78.87	80.81 93.93
+CPO	71.16 97.13	89.04 94.39	72.75 90.66	75.09 93.50	81.81 88.60	81.39 93.91
+StarPO	74.64 96.91	91.48 92.93	**78.07** 93.79	86.23 95.70	**82.21** 95.06	87.47 94.25
Qwen3-4B-Instruct						
Base	54.98 95.62	76.23 77.27	28.15 94.85	84.42 87.47	69.53 89.53	55.95 91.06
+SFT	77.63 89.92	89.55 92.07	43.49 85.10	81.99 84.73	71.58 91.98	81.56 97.63
+CPO	78.29 89.87	**90.11** 92.48	67.10 87.67	87.16 87.45	72.77 82.37	76.54 85.71
+StarPO	**83.74** 89.82	89.65 93.11	77.04 95.45	87.47 88.31	76.87 87.04	83.18 97.20
Other Systems						
Tower+	62.68 90.83	84.39 93.52	72.87 90.37	**86.46** 94.01	69.10 87.47	**91.44** 96.21
GPT-4o	65.09 90.06	82.74 90.04	70.31 80.81	72.73 87.23	77.58 89.84	66.88 93.29
Deepseek	64.00 89.87	66.57 82.72	69.84 90.24	64.52 93.30	67.78 92.74	62.72 90.84

Table 14: STAR scores on Guofeng test set. In each cell, the top value is the **Strict** score and the bottom value (in gray) is the **Relaxed** score.

System	Ideal	Structural Deviations		
	1-to-1	1-to-0	0-to-1	Other
LLAMA-3.1-8B-INSTRUCT				
Base	92.59	2.08	1.17	4.16
+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 15: Calculating STAR scores in Chinese ⇒ English language direction on News-Commentary test set by Gemini-2.5-Flash using prompts in Figure 5.

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 16: Performance in dCOMET 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 17: Comparison of different offline preference optimization algorithms.