M^{2}PO: Multi-Perspective Multi-Pair Preference Optimization for Machine Translation

Hao Wang[1], **Linlong Xu**[1], **Heng Liu**[1], **Yangyang Liu**[1], **Xiaohu Zhao**[1], **Bo Zeng**[1], **Liangying Shao**[1], **Yichen Dong**[1], **Xinwei Wu**[1], **Jiang Zhou**[1], **Tianyu Dong**[1], **Xiangxiang Zeng**[2], **Longyue Wang**[1], **Weihua Luo**[1]
[1]Alibaba Group, [2]Hunan University
huaiyu.wh@alibaba-inc.com

Abstract

Aligning Large Language Models (LLMs) to human preferences is pivotal for Machine Translation (MT), yet current approaches are often hindered by misleading reward signals. Our analysis reveals that prevailing Quality Estimation (QE) models exhibit a systematic blind spot towards **partial errors**—specifically partial hallucinations and omissions—often favoring superficially fluent but unfaithful translations. To address this, we propose **M^{2}PO** (**M**ulti-Perspective **M**ulti-Pair **P**reference **O**ptimization), a data-centric framework for preference optimization in machine translation. First, to correct the bias towards fluency, M^{2}PO uses a multi-perspective alignment mechanism that decouples semantic fidelity from fluency, prioritizing faithfulness via a curriculum strategy. Second, with the bias corrected, partial errors fall between perfect and severely incorrect translations, making them inefficient to learn via standard best-versus-worst comparisons. We thus introduce a multi-pair objective that leverages the full candidate list to capture these fine-grained error signals. Experiments on WMT23, WMT24, and FLORES-200 show that M^{2}PO enables a 9B model to outperform leading open-source baselines and achieve parity with proprietary models like GPT-4o and Gemini-2.0-Flash, demonstrating significant potential for efficient, high-fidelity LLM-based translation[1].

1 Introduction

The field of Machine Translation (MT) has been revitalized by large language models (LLMs) (Achiam et al., 2023; Chen et al., 2025; Lyu et al., 2025). While Supervised Fine-Tuning (SFT) has achieved remarkable success (Jiang et al., 2024; Dong et al., 2025), its reliance on Maximum Likelihood Estimation (MLE) often fails to capture the nuances of human preference (Xu et al., 2024a; Lu et al., 2024; Jiang et al., 2025). Although Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022) addresses this by aligning models with human values, its dependency on complex online interactions and training instability has shifted attention toward offline alternatives (She et al., 2024; Yang et al., 2025c; Tan and Monz, 2025). Consequently, direct preference-based learning methods, such as Direct Preference Optimization (DPO) (Rafailov et al., 2023) and Contrastive Preference Optimization (CPO) (Xu et al., 2024b), efficiently align models with "chosen" versus "rejected" translations.

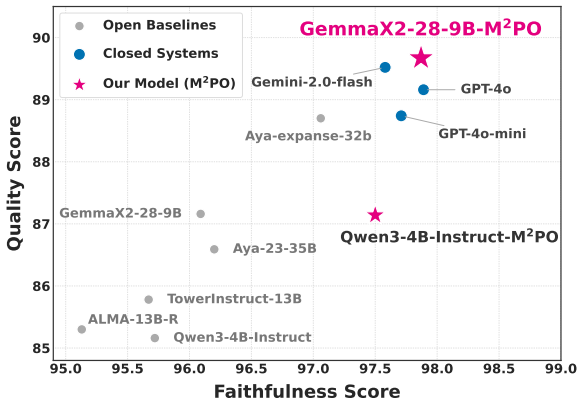

Figure 1: Translation Quality (XCOMET) vs. Faithfulness (Coverage Score (Wu et al., 2024)) averaged across 6 translation directions on the WMT23 benchmark.

Applying preference optimization to MT is constrained by reward reliability. Neural **Quality Estimation (QE)** models (Rei et al., 2023; Guerreiro et al., 2024) have replaced n-gram metrics (e.g., BLEU) as the preferred reward proxy due to their superior alignment with human perception. However, corroborating findings on the limitations of QE metrics (Deutsch et al., 2022; Dale et al., 2023a), our analysis (Section 3) reveals a critical bias: QE models often over-prioritize surface-level well-formedness. Consequently, they fail to

[1]Our code is available at https://github.com/AIDC-AI/Marco-MT/tree/master/MMPO.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 10315–10336

strictly penalize **partial errors** (exemplified in Appendix A), assigning deceptively high scores to candidates that are superficially fluent but semantically flawed. Compounding this, we find these errors concentrate in the *intermediate faithfulness range* (Figure 2). Crucially, even under a faithful ranking, this specific distribution renders them invisible to standard alignment strategies relying on single best-versus-worst pairs, which typically discard such non-extreme samples (He et al., 2024; Zeng et al., 2024; Sun et al., 2025).

To address these challenges, we introduce **M^{2}PO** (**M**ulti-Perspective **M**ulti-Pair **P**reference **O**ptimization), a data-centric framework designed to refine preference learning. Our framework consists of two complementary components. First, we employ *Multi-Perspective Preference Construction* to mitigate the limitations of QE metrics by integrating a discrete faithfulness coefficient. This mechanism strictly penalizes translation errors while preserving general quality, further enhanced by a dynamic curriculum that fuses expert scores with the model's evolving confidence. Second, we propose *Multi-Pair Joint Optimization* to maximize data utility by extending the standard pairwise objective to a dynamic multi-pair formulation. Through hierarchical contrasts, this strategy captures subtle errors, stabilized by a global ranking loss to regularize the model's output distribution.

We empirically validate M^{2}PO through extensive experiments on the WMT23, WMT24, and FLORES-200 benchmarks across six translation directions, applying our framework to two distinct base models (Qwen3-4B-Instruct (Yang et al., 2025a) and GemmaX2-28-9B (Cui et al., 2025)). As illustrated in Figure 1, M^{2}PO effectively propels open-source models to a new performance frontier, achieving high faithfulness and quality simultaneously. Most notably, despite its modest parameter scale, our **GemmaX2-28-9B-M^{2}PO** not only surpasses its data generator (GPT-4o-mini (Menick et al., 2024)) and significantly larger open-source baselines (e.g., 30B+), but also demonstrates capabilities comparable to—and in some cases outperforming—powerful proprietary systems like GPT-4o (Hurst et al., 2024) and Gemini-2.0-Flash (Comanici et al., 2025). We summarize our main contributions as follows:

- **Blind Spots in QE Rewards:** We identify a critical flaw in current QE metrics: they often prioritize fluency over faithfulness, providing deceptive signals that lead to reward hacking.
- **The M^{2}PO Framework:** We propose M^{2}PO, fusing multi-perspective alignment with multi-pair optimization. Diverging from standard pairwise methods, it leverages the full quality spectrum for maximal data efficiency.
- **Closing the Proprietary Gap:** M^{2}PO enables a 9B model to rectify supervision noise, surpassing its data generator (GPT-4o-mini) and matching proprietary systems (GPT-4o, Gemini-2.0-Flash).

2 Related Work

2.1 LLM-Based Translation and Alignment

While SFT builds foundational capabilities (Rei et al., 2025; Wu et al., 2025), alignment is pivotal for fine-grained quality. RLHF methodologies (Ouyang et al., 2022) diverge into online tracks like PPO (Schulman et al., 2017), GRPO (DeepSeek-AI, 2025), and GSPO (Zheng et al., 2025), which maximize rewards via iterative interactions (Ramos et al., 2024; Feng et al., 2025a; He et al., 2025; Li et al., 2025), and cost-effective offline implicit variants (Aakanksha et al., 2024; Yang et al., 2025b; Feng et al., 2025b). Recent advances in implicit optimization primarily focus on two parallel dimensions: signal construction and data efficiency. For the former, studies integrate QE metrics to scale supervision signals (Xu et al., 2025; He et al., 2024; Agrawal et al., 2024). In parallel, to address data efficiency, Set-MPO (Gupta et al., 2025) and LiPO (Liu et al., 2025) exploit the candidate spectrum via set-level contrasts and listwise ranking, respectively. However, their direct application to MT relies on the assumption of reliable rewards. Consequently, they remain susceptible to the *deceptive reward signals*, where fluent yet unfaithful translations mislead metrics.

2.2 Hallucination and Omission Mitigation

Ensuring faithfulness remains a central challenge in LLM-based MT, primarily manifesting through two distinct failure modes (Zhang et al., 2024). **Hallucinations** involve the fabrication of content not present in the source (Guerreiro et al., 2023; Himmi et al., 2024; Gogoulou et al., 2025), whereas **omissions** fail to convey critical source information despite the output's grammatical correctness (Yang et al., 2019; Vamvas and Sennrich, 2022; Dale et al., 2023b). Both errors are deceptive, as they are often masked by the model's high linguistic flu-

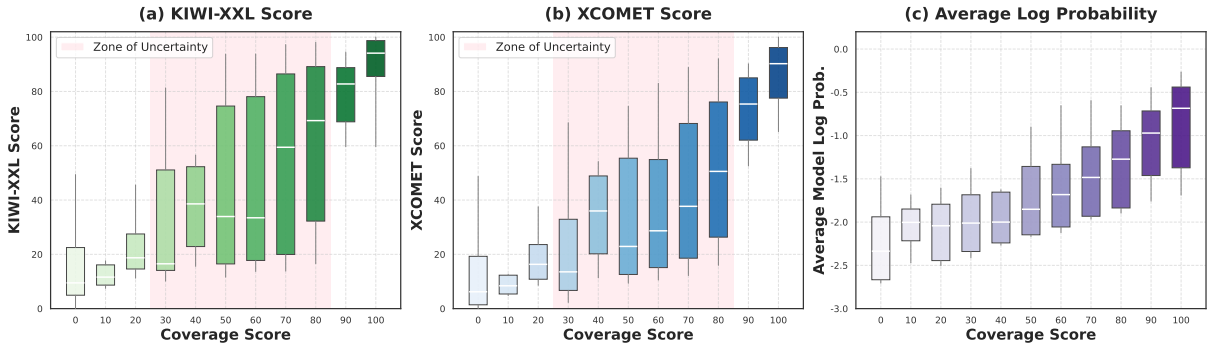

Figure 2: Analysis of **Metric Stability** across coverage levels (faithfulness) on the HalOmi benchmark. The plots illustrate the relationship between the Coverage Score and: (a) KIWI-XXL Score, (b) XCOMET Score, and (c) Sequence-level Average Model Log Probability (derived from the base model GemmaX2-28-9B). The shaded area highlights a "Zone of Uncertainty" (coverage range 30−80), where partial errors are prevalent.

ency. To address unfaithfulness, recent works like WAP (Wu et al., 2024) leverage word alignment signals as constraints; however, such approaches often compromise general translation quality (e.g., lower COMET scores) in exchange for stricter alignment.

3 The "Zone of Uncertainty": QE Instability vs. Model Consistency

While QE models serve as scalable proxies for human judgment, they often over-prioritize surface-level fluency (Deutsch et al., 2022), resulting in a systematic blind spot for nuanced errors—specifically, partial hallucinations and omissions, hereafter termed **partial errors**. To validate this, we analyze metric behavior on the HalOmi benchmark (Dale et al., 2023b). We contrast standard QE metrics, XCOMET-XXL[2] and COMET-KIWI-XXL[3] (hereafter **XCOMET** and **KIWI-XXL**), against **Coverage Score**—an LLM-based metric we implement using Gemini-2.0-Flash following the methodology of Wu et al. (2024), which quantifies the proportion of source semantic units preserved in the translation. Consistent with prior findings of Wu et al. (2024), our analysis (Appendix B) further validates Coverage Score as a reliable reference for faithfulness in this study.

The "Zone of Uncertainty" in QE Metrics. Figure 2 illustrates a sharp contrast in metric stability. Standard QE metrics effectively secure the boundaries, showing high consistency in both the Low-Score Region (overt failures) and High-Score Region (near-perfect translations). However, they falter in the critical interval (30−80)—denoted as the **"Zone of Uncertainty."** In this zone (dominated by partial errors), metrics lose their monotonic alignment with faithfulness (Panels a–b), frequently assigning misleadingly high scores to flawed candidates—echoing the critical blind spot for fluent hallucinations identified by Dale et al. (2023a). Detailed analysis in Appendix B corroborates that while QE excels at the extremes, it lacks the granularity to differentiate subtle failures within this unstable middle ground.

Model Confidence as a Stabilizer. Inspired by Dale et al. (2023a), who posit that model internal states outperform external metrics in detecting hallucinations, we validate this hypothesis on the HalOmi benchmark. In contrast to the volatility of QE, Figure 2(c) reveals that the sequence-level average log probability of the base model (GemmaX2-28-9B) maintains a robust monotonic trajectory across the coverage spectrum. Crucially, even when unfaithful translations appear linguistically fluent, the model's internal probability distribution reveals underlying uncertainty ("lying with hesitation"). This empirical evidence confirms that the model's internal belief can serve as a reliable regularizer against external reward blind spots, a key insight driving our M^{2}PO framework.

4 The M^{2}PO Framework

We introduce **M^{2}PO**, a data-centric framework addressing the reliability blind spots of QE metrics and the inefficiency of standard alignment. As illustrated in Figure 3, the framework is organized into two primary logical components structuring our discussion: *Multi-Perspective Preference Construction* (covering Stages 1 & 2) and *Multi-Pair Joint Optimization* (Stage 3).

[2] https://huggingface.co/Unbabel/XCOMET-XXL
[3] https://huggingface.co/Unbabel/wmt23-cometkiwi-da-xxl

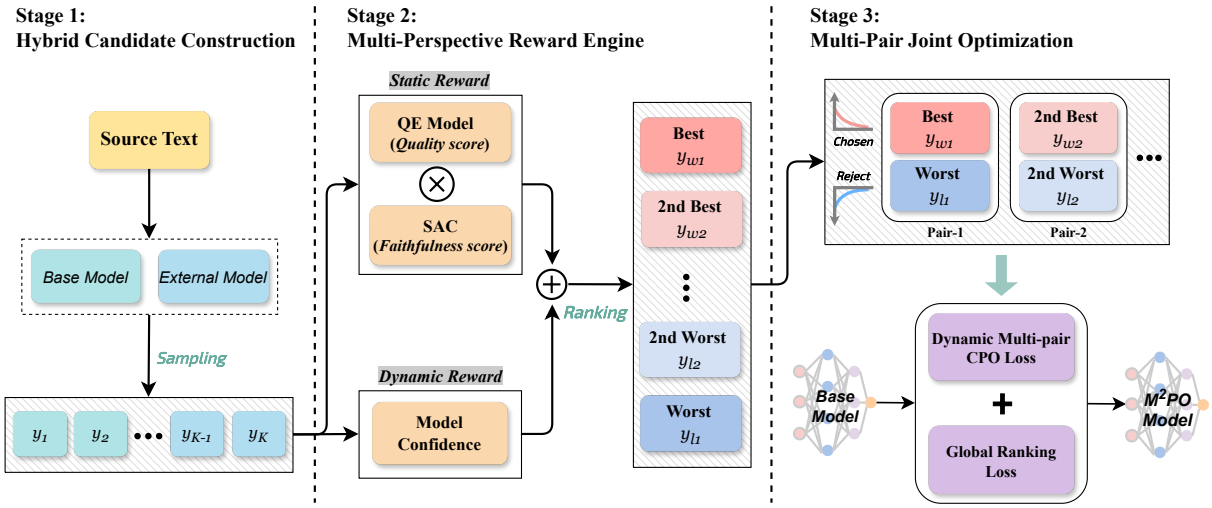

Figure 3: The M^{2}PO framework. The pipeline comprises: (1) Hybrid Candidate Construction for diverse sampling; (2) Multi-Perspective Reward Engine which ranks candidates by fusing static and dynamic signals; and (3) Multi-Pair Joint Optimization which updates the model using multiple contrastive pairs and a global ranking constraint.

4.1 Preliminaries

Preference Optimization Backbone. Let $\mathcal{D} = \{(x, y_w, y_l)\}$ denote a preference dataset, where x represents the source input, and y_w, y_l are the preferred and rejected candidate translations. To mitigate the memory overhead of reference-model-dependent methods like DPO, we optimize the policy π_θ using the reference-model-free **CPO** paradigm, which enforces fidelity by synergizing preference alignment with generation stability:

$$\mathcal{L}_{\text{CPO}}(x, y_w, y_l) = \underbrace{-\log \sigma\Big(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_\theta(y_l|x)}\Big)}_{\mathcal{L}_{\text{Pref}}} + \underbrace{\big(-\log \pi_\theta(y_w|x)\big)}_{\mathcal{L}_{\text{NLL}}} \tag{1}$$

Here, $\mathcal{L}_{\text{Pref}}$ maximizes the likelihood margin between y_w and y_l scaled by β. Crucially, the negative log-likelihood term $\mathcal{L}_{\text{NLL}}$ acts as a regularizer, anchoring the model to high-quality responses to prevent the degradation of linguistic fluency.

4.2 Multi-Perspective Preference Construction

We construct robust signals by integrating diverse sampling with multi-perspective scoring to capture the full spectrum of translation quality.

Hybrid Candidate Generation. To balance high-quality supervision with diverse error exposure, we construct a hybrid candidate pool $Y = \{y_1, \ldots, y_K\}$ of even size K for each source x. We sample $K/2$ candidates from a strong external model and the remaining $K/2$ from the base model. By uniformly applying high-randomness decoding ($T = 0.8$, Top-$p = 0.8$) to both sources, we aggressively diversify the candidate space. The resulting quality spectrum—ranging from strict fidelity to fluent yet unfaithful partial errors—provides the critical contrast required for learning fine-grained boundaries based solely on semantic alignment.

Calibrating QE with Semantic Alignment Classifier. Reliance on standard QE metrics for reward scoring often yields misleading signals within the "Zone of Uncertainty." To address this limitation, we prioritize LLM-based signals over traditional word alignment (e.g., WSPAlign (Wu et al., 2023)), as our analysis (Appendix B) highlights the latter's inefficiency in capturing nuanced partial errors. We introduce the *Semantic Alignment Classifier* (**SAC**), an LLM-driven discriminator that leverages discrete diagnostics to robustly rectify QE biases. SAC maps candidates into four tiers: *Full Match* (1.0), *High Partial* (0.7), *Low Partial* (0.3), and *No Match* (0.1), where the lowest tier serves as a safety floor to prevent signal erasure. Implementation details and sensitivity analyses are provided in Appendix C.2. We formulate the static reward r_s via multiplicative gating:

$$r_s(x, y_k) = r_{qe}(x, y_k) \cdot S_{align}(x, y_k) \tag{2}$$

where $r_{qe} \in [0, 100]$ is the KIWI-XXL score and S_{align} is the corresponding SAC coefficient. This formulation acts as a severity-aware veto, suppressing high-confidence yet unfaithful candidates (the "Zone of Uncertainty") towards the safety floor to ensure strict penalization and stability.

Figure 4: Visual comparison of Multi-Pair optimization (Zh→En). **Pair-1** shows a wide margin driven by a severe error, while **Pair-2** presents a "hard negative" (partial error) with a narrower margin. The optimization extends to **Pair-M**, corresponding to the $K/2$ split.

Dual-Perspective Fusion via Dynamic Curriculum. We construct a holistic reward by fusing the static faithfulness score $r_s(x, y_k)$ (Eq. 2) with the model's dynamic intrinsic confidence $r_d(x, y_k) = \frac{1}{|y_k|} \log \pi_\theta(y_k|x)$. To harmonize the bounded r_s and unbounded r_d, we apply z-score normalization over the candidate set Y via $\hat{r} = (r - \mu_Y)/\sigma_Y$. These standardized terms are integrated via a dynamic weighting scheme:

$$r_{\text{fused}}(x, y_k) = \sigma\left((1 - \alpha_t)\hat{r}_s + \alpha_t \hat{r}_d\right) \quad (3)$$

where $\sigma(\cdot)$ denotes the sigmoid function used for normalization. We adopt a linear curriculum by increasing $\alpha_t \in [0.1, 0.9]$ over the training steps t. This transitions the optimization from external grounding toward self-refinement, with the 0.9 upper bound preserving sufficient external supervision to prevent reward over-optimization.

4.3 Multi-Pair Joint Optimization

To maximize data efficiency, we propose a Dynamic Multi-Pair Strategy that transcends standard best-vs-worst comparisons. As illustrated in Figure 4, while traditional methods rely on the sharpest distinction (Pair-1), they neglect the intermediate "Zone of Uncertainty." M^{2}PO addresses this by mining the full candidate spectrum. We sort candidates by descending r_{fused} to obtain the ranked set $\{y_{(1)}, \ldots, y_{(K)}\}$ and construct $M = K/2$ pairs by coupling the top-i and bottom-i candidates: $(y_w^i, y_l^i) = (y_{(i)}, y_{(K-i+1)})$. This hierarchical coupling creates a gradient of difficulty, ranging from clear quality gaps to subtle "hard negatives" (e.g., Pair-2) that demand deeper discrimination.

Local: Dynamic Multi-Pair CPO. Building upon these hierarchical contrasts, we minimize a joint preference-NLL objective, weighting each pair by its normalized margin gap $\Delta r^i = r_{\text{fused}}(x, y_w^i) - r_{\text{fused}}(x, y_l^i)$:

$$\mathcal{L}_{\text{DM-CPO}} = \sum_{i=1}^{M} \left(\frac{\Delta r^i}{\sum_{j=1}^{M} \Delta r^j} \right) \cdot \mathcal{L}_{\text{CPO}}(x, y_w^i, y_l^i) \quad (4)$$

This weighting anchors stability by emphasizing distinct quality gaps, while attenuating ambiguous pairs to prevent overfitting in the "Zone of Uncertainty."

Global: Ranking Regularization. To complement local pairwise discrimination, we enforce global ordinal consistency via a Listwise Ranking Loss (Cao et al., 2007):

$$\mathcal{L}_{\text{Rank}} = -\sum_{k=1}^{K} P_{\text{fused}}(y_k|x) \log P_\theta(y_k|x) \quad (5)$$

where the target distribution $P_{\text{fused}}(y_k|x)$ is defined as $\text{softmax}(r_{\text{fused}}(x, y_k)/\tau)$, and the model distribution $P_\theta(y_k|x)$ is derived by re-normalizing the policy log-probabilities $\log \pi_\theta(y_k|x)$ over the candidate set Y via softmax.

Joint Objective. The final objective sums these components with weighting hyperparameter λ_{rank}:

$$\mathcal{L}_{\text{M}^2\text{PO}} = \mathcal{L}_{\text{DM-CPO}} + \lambda_{\text{rank}} \mathcal{L}_{\text{Rank}} \quad (6)$$

5 Experiments

5.1 Datasets and Baselines

Training Data: M^{2}PO-Prefer. We construct our training set, M^{2}PO-Prefer, using seeds derived from WMT22 (Kocmi et al., 2022), IWSLT (Cettolo et al., 2017), and FLORES-200-dev (Costa-Jussà et al., 2022). Following §4.2, we construct the pool ($K = 8$) with 4 candidates each from GPT-4o-mini (external model) and GemmaX2-28-9B (base model). After rigorous filtering, the dataset comprises ∼20,000 source inputs (totaling ∼160,000 candidates) across six directions (En↔{Zh, De, Ja}), demonstrating remarkable cost-efficiency (see analysis in Appendix C.3). Detailed dataset statistics are in Appendix D.

Test Benchmarks. We evaluate on official benchmarks strictly disjoint from training seeds: WMT23 (Kocmi et al., 2023) and FLORES-200-test (Costa-Jussà et al., 2022) for bidirectional evaluation (En↔X), and WMT24 (Kocmi et al., 2024) for unidirectional En→X assessment.

Model	AVG	En→Zh	En→De	En→Ja	Zh→En	De→En	Ja→En
Gemini-2.0-Flash	97.58 / 89.52	97.25 / 88.99	97.94 / **89.13**	97.65 / 91.09	97.67 / 93.15	97.95 / 86.66	97.01 / 88.12
GPT-4o	**97.89** / 89.16	97.36 / 87.96	**98.75** / 88.20	97.84 / 90.25	98.09 / 92.96	98.08 / 87.05	97.20 / 88.55
GPT-4o-mini	97.71 / 88.74	97.40 / 87.19	98.50 / 86.88	97.95 / 90.15	97.85 / 93.09	97.72 / 86.80	96.85 / 88.35
Aya-expanse-32B	97.06 / 88.70	96.96 / 88.50	97.71 / 86.51	97.33 / 90.75	97.03 / 93.34	97.09 / 85.93	96.21 / 87.18
Aya-23-35B	96.20 / 86.59	96.03 / 86.67	97.52 / 84.62	96.32 / 88.76	96.08 / 90.11	96.84 / 84.73	94.39 / 84.65
TowerInstruct-13B	95.67 / 85.78	96.17 / 86.99	97.20 / 83.90	94.97 / 81.56	96.22 / 91.66	97.17 / 85.90	92.28 / 84.64
ALMA-13B-R	95.13 / 85.30	95.63 / 86.63	96.85 / 83.97	93.99 / 79.84	95.79 / 92.12	96.92 / 85.55	91.60 / 83.70
Qwen3-4B-Instruct	95.72 / 85.16	95.88 / 85.65	95.77 / 79.76	94.96 / 85.70	96.67 / 91.25	96.88 / 83.81	94.17 / 84.77
+ SFT	96.76 / 86.20	97.02 / 87.34	96.86 / 80.13	95.49 / 86.83	97.47 / 92.43	97.97 / 84.62	95.76 / 85.82
+ CPO	96.94 / 86.57	97.05 / 87.81	97.08 / 80.79	96.07 / 87.12	97.53 / 92.67	98.02 / 84.79	95.86 / 86.23
+ M²PO	97.50 / 87.14	97.23 / 88.84	97.76 / 81.27	97.13 / 87.72	**98.25** / 92.89	**98.43** / 85.12	96.17 / 86.98
GemmaX2-28-9B	96.09 / 87.16	96.02 / 87.59	96.89 / 85.21	96.61 / 87.86	96.37 / 91.60	96.36 / 85.70	94.26 / 85.01
+ SFT	96.81 / 87.79	96.64 / 88.18	97.49 / 85.47	97.40 / 88.84	97.07 / 91.94	97.34 / 86.32	94.93 / 86.00
+ Set-MPO†	96.90 / 88.06	96.48 / 87.99	97.28 / 85.67	97.75 / 89.12	97.19 / 92.26	96.98 / 86.20	95.73 / 87.11
+ LiPO†	97.08 / 87.97	97.15 / 88.01	97.33 / 85.89	97.69 / 88.90	97.37 / 91.84	97.06 / 86.31	95.88 / 86.88
+ CPO	97.13 / 88.59	96.98 / 88.28	97.68 / 85.51	97.62 / 90.20	97.23 / 92.89	97.45 / 86.61	95.83 / 88.02
+ M²PO	97.87 / **89.67**	**97.55** / **89.65**	98.13 / 86.21	**97.96** / **92.16**	97.98 / **93.46**	98.41 / **87.56**	**97.21** / **88.98**

Table 1: Main results on the WMT23 benchmark. Results are reported in the format of **Coverage / XCOMET**. Model names are color-coded by type: Proprietary, Open-source baselines, and Our experiments. **Bold** indicates the best result overall. Colored background indicates the best result among open-source models. † indicates methods originally proposed for general tasks, which we adapted and reproduced for machine translation.

Comparison Systems. We benchmark M²PO against three groups: (1) **Proprietary Systems**: Gemini-2.0-Flash, GPT-4o, and GPT-4o-mini;[4] (2) **Open-source LLMs**: Including general-purpose models (Aya-expanse-32B, Aya-23-35B) (Dang et al., 2024; Aryabumi et al., 2024) and translation-specialized models (TowerInstruct-13B, ALMA-13B-R) (Alves et al., 2024; Xu et al., 2024b); and (3) **Controlled Baselines**: Implementation of SFT (trained on the top-ranked candidates from M²PO-Prefer) and preference-based methods (CPO, Set-MPO, LiPO) utilizing standard KIWI-XXL as the reward signal. Supplementary ablations on SFT gold references and CPO reward configurations are provided in Appendix E and §6.5, respectively.

5.2 Implementation and Evaluation

Model Configuration. We employ two advanced open-source base models: **GemmaX2-28-9B**[5] and **Qwen3-4B-Instruct**[6]. Both are adapted via LoRA (Hu et al., 2022) ($r = 32, \alpha = 64$) and optimized using AdamW (Loshchilov and Hutter, 2019) with a learning rate of 5×10^{-5}, batch size of 32, and sequence length of 512, utilizing a 0.1 warmup ratio and 1.0 gradient clipping. For M²PO, we set $\beta = 0.1$, $\lambda_{\text{rank}} = 0.5$, and $\tau = 1.0$. Our experiments run on an NVIDIA H100 GPU for 2 epochs, requiring approximately 2 hours.

Evaluation Protocol. Inference is conducted via vLLM (Kwon et al., 2023) ($T = 0.3$, Top-$p = 0.3$). We adopt a comprehensive dual-aspect metric suite:

(1) Faithfulness: We employ the **Coverage Score** (via Gemini-2.0-Flash (Wu et al., 2024)) as our primary faithfulness metric, supported by the discussion in §3.

(2) Translation Quality: We employ the widely recognized **XCOMET** for reference-free evaluation. While we acknowledge QE blind spots for partial errors, our comprehensive analysis (§3 and Appendix B) confirms that XCOMET remains a highly reliable discriminator *within the high-faithfulness regime*. Thus, once faithfulness is secured, it effectively measures fluency and nuance. To mitigate potential bias, we complement this with reference-based **COMET-22**[7] (Rei et al., 2020).

5.3 Main Results

Table 1 presents the primary performance on WMT23, where M²PO achieves leading results among open baselines, effectively closing the gap with proprietary systems. Specifically, our

[4] Versions: `gemini-2.0-flash-001`, `gpt-4o-2024-08-06`, and `gpt-4o-mini-2024-07-18`.

[5] https://huggingface.co/ModelSpace/GemmaX2-28-9B-v0.1

[6] https://huggingface.co/Qwen/Qwen3-4B-Instruct-2507

[7] https://huggingface.co/Unbabel/wmt22-comet-da

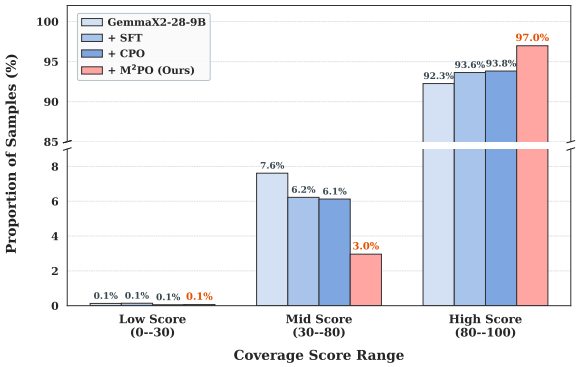

Figure 5: Distribution of sample proportions across Low (0–30), Mid (30–80), and High (80–100) Coverage Score intervals on the WMT23 benchmark.

GemmaX2-28-9B-M²PO outperforms all open baselines—including the significantly larger Aya-expanse-32B—by achieving a 97.87 Coverage Score and 89.67 XCOMET. Notably, it surpasses its data generator (GPT-4o-mini) and even exceeds GPT-4o in translation quality (+0.51 XCOMET) while maintaining comparable faithfulness. At smaller scales, *Qwen3-4B-Instruct-M²PO* demonstrates exceptional efficiency, surpassing 13B baselines and securing the highest Coverage on Zh→En (98.25) and De→En (98.43) among all systems.

Controlled comparisons further confirm the algorithmic superiority of our framework. M²PO yields substantial gains over standard CPO (+1.08 XCOMET) as well as advanced set-level (SetMPO) and listwise (LiPO) adaptations. This validates that the synergy of multi-perspective preference labeling and multi-pair optimization provides cleaner, more granular learning signals than general-purpose alignment methods.

Beyond standard benchmarks, we rigorously validate robustness (Appendix F). First, consistent effectiveness on COMET-22 confirms that gains are not artifacts of specific metric optimization. Second, FLORES-200 and WMT24 evaluations verify generalization across diverse domains and temporal shifts, ruling out overfitting. Finally, qualitative case studies (Appendix A) provide concrete evidence of M²PO rectifying subtle hallucinations and omissions that persist in baselines.

6 Analysis

6.1 Mitigating Partial Errors in the "Zone of Uncertainty"

Figure 5 visualizes the full score distribution, situating the intermediate "Zone of Uncertainty" (30–80) between the low and high extremes. A defining challenge in this regime is **sparsity**: deceptive partial errors comprise only (∼6–7%) of the data. Instead of actively correcting them, standard optimization objectives tend to prioritize the majority of easier patterns, effectively diluting the learning signal from these rare, fine-grained failures. Consequently, standard methods fail to mine these subtle signals; SFT and CPO plateau at comparable error rates (6.2% and 6.1%), offering limited gains over the base model. In contrast, M²PO explicitly amplifies these sparse instances via faithfulness-weighted penalties. This acts as a precision filter, reducing partial errors by ∼51% (to 3.0%) and propelling the distribution into the high-faithfulness regime (>80 coverage), peaking at 97.0%. This confirms M²PO rectifies fine-grained failures evading conventional alignment.

System	Quality Score			Average Metric	
	Zh	De	Ja	Quality	Faith. (%)
GPT-4o-mini	4.11	**4.32**	4.56	4.33	87.11
GemmaX2-28-9B	3.73	3.83	3.04	3.53	72.67
M²PO (Ours)	**4.14**	4.30	**4.67**	**4.37**	**89.11**

Table 2: Human evaluation results on 450 sampled WMT23 instances (150 per direction across En-Zh, En-De, En-Ja). **Quality** measures holistic fluency and adequacy (0–5); **Faith.** denotes the percentage of "Full Match" labels averaged across languages.

6.2 Human Evaluation Validation

We conducted a blind expert evaluation on 450 representative WMT23 instances. To ensure balanced coverage of challenging cases, we employed stratified sampling across the "Zone of Uncertainty" and remaining regions. Annotators adhered to a dual-aspect protocol: (1) holistic Quality Ratings (0–5) adapted from WMT22 standards (Kocmi et al., 2022); and (2) manual Faithfulness Classification via our fine-grained SAC taxonomy.

As shown in Table 2, M²PO demonstrates superior performance across both dimensions. It achieves peak faithfulness (89.11%), marking a substantial +16.4% gain over the base model. Notably, it surpasses its data generator, GPT-4o-mini (87.11%), acting as a *denoising filter* to rectify subtle errors. In addition to exceptional faithfulness, M²PO attains the highest holistic Quality Score (4.37), outperforming the proprietary baseline (4.33) and excelling in linguistically distant languages (e.g., 4.67 in En-Ja), thereby effectively

Base Objective	Standard	+ M²PO (Ours)
Base Model	96.09 / 87.16	—
DPO	96.91 / 88.48	97.54 / 89.15
KTO	96.71 / 88.39	97.37 / 89.43
SimPO	97.06 / **88.74**	97.73 / 89.55
ORPO	96.47 / 87.86	96.85 / 88.41
CPO	**97.13** / 88.59	**97.87 / 89.67**

Table 3: Generalization across different preference objectives on WMT23. "+ M²PO" indicates plugging the standard loss (e.g., $\mathcal{L}_{\text{DPO}}$) into our framework by replacing $\mathcal{L}_{\text{CPO}}$ in Eq. 4. Scores are **Coverage / XCOMET**.

Reward Configuration for M²PO	Coverage
Base (GemmaX2-28-9B)	96.09
MetricX-24-XXL only	97.35
MetricX-24-XXL + SAC	97.72
KIWI-XXL only	96.99
KIWI-XXL + SAC (Ours)	**97.87**

Table 4: Performance comparison on WMT23 across different reward proxies. We evaluate the impact of the SAC penalty when integrated with KIWI-XXL and MetricX-24-XXL.

reconciling the trade-off between faithfulness and fluency.

6.3 Generalization Across Preference Objectives

M²PO is a flexible framework designed to enhance offline preference optimization. We validate its generality by extending it to diverse DPO-style objectives: DPO (Rafailov et al., 2023), KTO (Ethayarajh et al., 2024), SimPO (Meng et al., 2024), and ORPO (Hong et al., 2024).

Integration Mechanism. The "+ M²PO" configuration represents a dual enhancement. First, we use our multi-perspective rewards to construct high-quality preference pairs. Second, we substitute the specific term $\mathcal{L}_{\text{CPO}}$ in Eq. 4 with the target objective (e.g., $\mathcal{L}_{\text{DPO}}$) while retaining the dynamic weights. This allows standard baselines to benefit from both our refined data selection strategy and our dynamic calibration mechanism.

Performance. Table 3 shows consistent gains across all baselines. Enhancing the strongest baseline (CPO) yields peak performance (97.87 Coverage), while KTO sees a substantial boost of +1.04 XCOMET. Furthermore, we demonstrate that M²PO remains competitive even against computation-heavy Online-RL algorithms (see Appendix G for detailed comparisons).

6.4 Generalization to Alternative Reward Proxies

To evaluate the generalization of M²PO, we substitute **KIWI-XXL** with **MetricX-24-XXL**[8], a leading regression-based evaluation metric. As shown in Table 4, standalone MetricX-24-XXL (97.35) already outperforms standalone KIWI-XXL (96.99). However, integrating our SAC penalty with MetricX-24-XXL yields further gains (+0.37 Coverage). While our default KIWI-XXL + SAC configuration marginally achieves the highest overall score (97.87), the broader takeaway is that integrating SAC consistently improves the evaluated continuous metrics, resulting in comparable alignment efficacy across the hybrid configurations.

To understand why explicit faithfulness supervision remains vital even with stronger continuous metrics, we conduct a qualitative analysis of partial errors (detailed in Appendix H). We observe that both MetricX-24-XXL and KIWI-XXL share a critical blind spot within the "Zone of Uncertainty" (introduced in §3): **Ranking Inconsistency**. They tend to prioritize surface-level fluency over strict semantic faithfulness, occasionally assigning better scores to highly fluent outputs with major omissions than to those with minor errors. SAC mitigates this vulnerability by providing a robust, discrete penalty that explicitly corrects these ranking inversions. This confirms that continuous QE proxies (assessing overall translation quality) and SAC (acting as a strict semantic guardrail) are fundamentally complementary.

6.5 Ablation Studies

We conduct systematic ablations on WMT23 (Table 5) to dissect M²PO. To rigorously attribute performance gains, we structure the analysis into two dimensions: validating the fundamental optimization objectives and isolating the efficacy of specific algorithmic mechanisms.

Impact of Optimization Objectives. The DM-CPO engine ($\mathcal{L}_{\text{DM-CPO}}$) acts as the cornerstone; its removal results in the sharpest decline (-1.11 Coverage and -1.58 XCOMET), confirming its primacy in enforcing semantic adherence. Complementarily,

[8] https://huggingface.co/google/metricx-24-hybrid-xxl-v2p6

Model Config.	Coverage	XCOMET
Full M²PO	**97.87**	**89.67**
Impact of Loss Components		
w/o $\mathcal{L}_{\text{DM-CPO}}$	96.76	88.09
w/o $\mathcal{L}_{\text{Rank}}$	97.15	89.19
Impact of Core Strategies		
w/o SAC Fidelity Filter	96.99	89.28
w/o Dynamic Fusion	97.06	89.24
w/o Multi-Pair	97.42	89.16
Base (GemmaX2-28-9B)	96.09	87.16

Table 5: Ablation studies on the WMT23 benchmark.

Model Configuration	Coverage	XCOMET
Base (GemmaX2-28-9B)	96.09	87.16
Open-Source Pipeline		
+ CPO	96.74	88.42
+ M²PO (w/o SAC)	96.83	88.73
+ M²PO (w/ Open-Source SAC)	**97.37**	**89.16**
Proprietary Pipeline		
+ CPO	97.13	88.59
+ M²PO (w/o SAC)	96.99	89.28
+ M²PO (w/ Proprietary SAC)	**97.87**	**89.67**

Table 6: Ablation study on WMT23 evaluating the fully open-source framework. The *Open-Source Pipeline* uses the base model for candidate generation and Gemma-3-27B-it for SAC labeling, while the *Proprietary Pipeline* uses GPT-4o-mini for both tasks.

excluding the Ranking Loss ($\mathcal{L}_{\text{Rank}}$) compromises global ordinal consistency, validating that global anchoring serves as a necessary safeguard against faithfulness degradation (see Appendix I for weight sensitivity analysis).

Impact of Core Strategies. We further validate specific algorithmic mechanisms. (1) SAC Fidelity Filter: Removing this constraint ($S_{align} = 1$) drops Coverage to 96.99, confirming it suppresses fluent errors typically overlooked by standard QE (sensitivity analysis in Appendix C.2). (2) Dynamic Fusion: Ablating the curriculum ($\alpha_t = 0$) impairs performance, underscoring the need to balance *external grounding* with *intrinsic confidence* to mitigate static proxy biases (sensitivity analysis in Appendix J). (3) Multi-Pair Contrast: Removing this component reduces our method to a standard single-pair CPO baseline trained with our augmented reward (KIWI-XXL + SAC + Model Confidence). The consistent performance drop across all metrics proves that contrasting multiple candidates provides a clearer learning signal than relying on a single positive-negative pair.

6.6 Independence from Proprietary Distillation

A critical question is whether M²PO's gains stem primarily from proprietary teacher distillation (e.g., GPT-4o-mini) or the proposed multi-pair optimization framework itself. To disentangle this, we compare an *Open-Source Pipeline* against a *Proprietary Pipeline*.

The *Open-Source Pipeline* restricts candidate generation to the base model (GemmaX2-28-9B). Because this base model is specialized for translation and lacks robust evaluation capabilities, we utilize Gemma-3-27B-it[9] for SAC labeling. Conversely, the *Proprietary Pipeline* relies entirely on GPT-4o-mini for both candidate generation and labeling.

As shown in Table 6, the open-source M²PO consistently outperforms the CPO baseline (+0.63 Coverage, +0.74 XCOMET). Furthermore, removing the SAC signal degrades performance (to 96.83 Coverage and 88.73 XCOMET), mirroring the ablation trends of the proprietary setting. These results confirm that M²PO's improvements are intrinsically driven by its multi-pair optimization and multi-reward signals, rather than simply acting as a distillation vehicle. Crucially, this validates a fully open-source paradigm for deploying M²PO without reliance on commercial APIs.

7 Conclusion

We present M²PO, a data-centric framework addressing the blind spots of standard reward models and the data inefficiency in MT preference optimization. By integrating a multi-perspective alignment curriculum with a multi-pair strategy augmented by listwise ranking, M²PO effectively targets the "Zone of Uncertainty," rectifying deceptive partial errors overlooked by conventional methods. Extensive evaluations across WMT23, WMT24, and FLORES-200 confirm the framework's efficacy on 4B and 9B models, showing consistent gains in both faithfulness and translation quality. Results highlight that precise preference alignment enables compact open-source models to rival larger proprietary systems, establishing M²PO as a robust pathway toward high-fidelity MT.

[9] https://huggingface.co/google/gemma-3-27b-it

8 Limitations

While M^{2}PO demonstrates strong performance on compact models (4B/9B) across major translation directions, validating its scalability to larger architectures (e.g., 70B+) and low-resource languages remains an important direction for future work. Regarding the data and evaluation pipeline, our primary experiments rely on proprietary models (GPT-4o-mini, Gemini-2.0-Flash) to provide high-fidelity training signals and robust assessment. Although this reliance on commercial APIs may limit the immediate reproducibility of the full workflow, we have taken initial steps toward mitigation by validating a fully open-source alternative (§6.6). Future work will further optimize this offline-deployable framework to narrow the gap with proprietary distillation and extend it to broader translation scenarios.

Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable and constructive feedback. This work was supported by Alibaba Group.

References

Aakanksha, Arash Ahmadian, Beyza Ermis, Seraphina Goldfarb-Tarrant, Julia Kreutzer, Marzieh Fadaee, and Sara Hooker. 2024. The multilingual alignment prism: Aligning global and local preferences to reduce harm. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA. Association for Computational Linguistics.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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 the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA. Association for Computational Linguistics.

Duarte M. Alves, José Pombal, Nuno Miguel Guerreiro, Pedro Henrique Martins, João Alves, M. 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.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan N. Gomez, Phil Blunsom, Marzieh Fadaee, and 2 others. 2024. Aya 23: Open weight releases to further multilingual progress. *CoRR*, abs/2405.15032.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: from pairwise approach to listwise approach. In *ICML*, volume 227 of *ACM International Conference Proceeding Series*, pages 129–136. ACM.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuhito Sudoh, Koichiro Yoshino, and Christian Federmann. 2017. Overview of the IWSLT 2017 evaluation campaign. In *Proceedings of the 14th International Conference on Spoken Language Translation*, Tokyo, Japan. International Workshop on Spoken Language Translation.

Andong Chen, Lianzhang Lou, Kehai Chen, Xuefeng Bai, Yang Xiang, Muyun Yang, Tiejun Zhao, and Min Zhang. 2025. Benchmarking LLMs for translating classical Chinese poetry: Evaluating adequacy, fluency, and elegance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Suzhou, China. Association for Computational Linguistics.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Menglong Cui, Pengzhi Gao, Wei Liu, Jian Luan, and Bin Wang. 2025. Multilingual machine translation with open large language models at practical scale: An empirical study. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Albuquerque, New Mexico. Association for Computational Linguistics.

David Dale, Elena Voita, Loic Barrault, and Marta R. Costa-jussà. 2023a. Detecting and mitigating hallucinations in machine translation: Model internal workings alone do well, sentence similarity Even better. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada. Association for Computational Linguistics.

David Dale, Elena Voita, Janice Lam, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Loïc Barrault, and Marta R. Costa-jussà. 2023b. HalOmi: A manually annotated benchmark for multilingual hallucination and omission detection in machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore. Association for Computational Linguistics.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, and 1 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *arXiv preprint arXiv:2412.04261*.

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

Daniel Deutsch, Rotem Dror, and Dan Roth. 2022. On the limitations of reference-free evaluations of generated text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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 the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vienna, Austria. Association for Computational Linguistics.

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

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 Association for Computational Linguistics: EMNLP 2025*, Suzhou, China. Association for Computational Linguistics.

Zhaopeng Feng, Jiahan Ren, Jiayuan Su, Jiamei Zheng, Hongwei Wang, and Zuozhu Liu. 2025b. MT-RewardTree: A comprehensive framework for advancing LLM-based machine translation via reward modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, Suzhou, China. Association for Computational Linguistics.

Evangelia Gogoulou, Shorouq Zahra, Liane Guillou, Luise Dürlich, and Joakim Nivre. 2025. Can LLMs detect intrinsic hallucinations in paraphrasing and machine translation? In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM^2)*, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

Nuno M. Guerreiro, Duarte M. Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023. Hallucinations in large multilingual translation models. *Transactions of the Association for Computational Linguistics*, 11.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12.

Taneesh Gupta, Rahul Madhavan, Xuchao Zhang, Nagarajan Natarajan, Chetan Bansal, and Saravan Rajmohan. 2025. Multi-preference optimization: Generalizing dpo via set-level contrasts. *Preprint*, arXiv:2412.04628.

Minggui He, Yilun Liu, Shimin Tao, Yuanchang Luo, Hongyong Zeng, Chang Su, Li Zhang, Hongxia Ma, Daimeng Wei, Weibin Meng, Hao Yang, Boxing Chen, and Osamu Yoshie. 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 the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Mexico City, Mexico. Association for Computational Linguistics.

Anas Himmi, Guillaume Staerman, Marine Picot, Pierre Colombo, and Nuno M Guerreiro. 2024. Enhanced hallucination detection in neural machine translation through simple detector aggregation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA. Association for Computational Linguistics.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. ORPO: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *ICLR*. OpenReview.net.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas,

Emma Bou Hanna, Florian Bressand, and 1 others. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Ruili Jiang, Kehai Chen, Xuefeng Bai, Zhixuan He, Juntao Li, Muyun Yang, Tiejun Zhao, Liqiang Nie, and Min Zhang. 2025. A survey on human preference learning for aligning large language models. *ACM Computing Surveys*, 58(6):1–39.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, and 3 others. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, and 3 others. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, Singapore. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *SOSP*, pages 611–626. ACM.

Tianjiao Li, Mengran Yu, Chenyu Shi, Yanjun Zhao, Xiaojing Liu, Qi Zhang, Xuanjing Huang, Qiang Zhang, and Jiayin Wang. 2025. RIVAL: Reinforcement learning with iterative and adversarial optimization for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, Suzhou, China. Association for Computational Linguistics.

Tianqi Liu, Zhen Qin, Junru Wu, Jiaming Shen, Misha Khalman, Rishabh Joshi, Yao Zhao, Mohammad Saleh, Simon Baumgartner, Jialu Liu, Peter J Liu, and Xuanhui Wang. 2025. LiPO: Listwise preference optimization through learning-to-rank. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Albuquerque, New Mexico. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *ICLR (Poster)*. OpenReview.net.

Yinquan Lu, Wenhao Zhu, Lei Li, Yu Qiao, and Fei Yuan. 2024. LLaMAX: Scaling linguistic horizons of LLM by enhancing translation capabilities beyond 100 languages. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Miami, Florida, USA. Association for Computational Linguistics.

Xinglin Lyu, Wei Tang, Yuang Li, Xiaofeng Zhao, Ming Zhu, Junhui Li, Yunfei Lu, Min Zhang, Daimeng Wei, Hao Yang, and Min Zhang. 2025. DoCIA: An online document-level context incorporation agent for speech translation. In *Findings of the Association for Computational Linguistics: ACL 2025*, Vienna, Austria. Association for Computational Linguistics.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235.

Jacob Menick, Kevin Lu, Shengjia Zhao, E Wallace, H Ren, H Hu, N Stathas, and F Petroski Such. 2024. Gpt-4o mini: advancing cost-efficient intelligence. *Open AI: San Francisco, CA, USA*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

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 *NeurIPS*.

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.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André F. T. Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of*

the Eighth Conference on Machine Translation, Singapore. Association for Computational Linguistics.

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

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online. Association for Computational Linguistics.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347.

Shuaijie She, Wei Zou, Shujian Huang, Wenhao Zhu, Xiang Liu, Xiang Geng, and Jiajun Chen. 2024. MAPO: Advancing multilingual reasoning through multilingual-alignment-as-preference optimization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Bangkok, Thailand. Association for Computational Linguistics.

Haoxiang Sun, Ruize Gao, Pei Zhang, Baosong Yang, and Rui Wang. 2025. Enhancing machine translation with self-supervised preference data. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vienna, Austria. Association for Computational Linguistics.

Shaomu Tan and Christof Monz. 2025. ReMedy: Learning machine translation evaluation from human preferences with reward modeling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Suzhou, China. Association for Computational Linguistics.

Jannis Vamvas and Rico Sennrich. 2022. As little as possible, as much as necessary: Detecting over- and undertranslations with contrastive conditioning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Dublin, Ireland. Association for Computational Linguistics.

Di Wu, Yibin Lei, and Christof Monz. 2025. Calibrating translation decoding with quality estimation on llms. *CoRR*, abs/2504.19044.

Qiyu Wu, Masaaki Nagata, Zhongtao Miao, and Yoshimasa Tsuruoka. 2024. Word alignment as preference for machine translation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA. Association for Computational Linguistics.

Qiyu Wu, Masaaki Nagata, and Yoshimasa Tsuruoka. 2023. WSPAlign: Word alignment pre-training via large-scale weakly supervised span prediction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada. Association for Computational Linguistics.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024a. A paradigm shift in machine translation: Boosting translation performance of large language models. In *ICLR*. OpenReview.net.

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 *ICLR*. OpenReview.net.

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

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Wen Yang, Junhong Wu, Chen Wang, Chengqing Zong, and Jiajun Zhang. 2025b. Implicit cross-lingual rewarding for efficient multilingual preference alignment. In *Findings of the Association for Computational Linguistics: ACL 2025*, Vienna, Austria. Association for Computational Linguistics.

Wen Yang, Junhong Wu, Chen Wang, Chengqing Zong, and Jiajun Zhang. 2025c. Language imbalance driven rewarding for multilingual self-improving. In *ICLR*. OpenReview.net.

Zonghan Yang, Yong Cheng, Yang Liu, and Maosong Sun. 2019. Reducing word omission errors in neural machine translation: A contrastive learning approach. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. Association for Computational Linguistics.

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

Hongbin Zhang, Kehai Chen, Xuefeng Bai, Yang Xiang, and Min Zhang. 2024. Paying more attention to source context: Mitigating unfaithful translations from large language model. In *Findings of the Association for Computational Linguistics: ACL 2024*, Bangkok, Thailand. Association for Computational Linguistics.

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.

System	Translation Output	Training		Testing	
		KIWI-XXL	SAC	XCOMET	Coverage
Source (zh→en): 口水鸡应该是熟的，但收到的是生肉，没办法吃					
GemmaX2-28-9B	The cold chicken should be cooked, but what I received was raw meat, making it inedible.	62.01	**High Partial**	85.57	85
GPT-4o-mini	The mouth-watering chicken should be cooked, but what I received was raw meat, which was inedible.	54.49	**Full match**	84.04	100
M^{2}PO	Poached chicken with chili sauce should be cooked, but what I received was raw meat, making it inedible.	63.48	**Full match**	89.52	100
Source (en→zh): At the time, nearly 100 residents were evacuated from the area.					
GemmaX2-28-9B	当时，近100名居民从该地区被疏 evacuated。	80.08	**Low Partial**	99.00	85
GPT-4o-mini	当时，几乎有100名居民从该地区撤离。	73.88	**Full match**	99.20	100
M^{2}PO	当时，近100名居民被从该地区疏散。	96.29	**Full match**	99.98	100
Source (de→en): Entsorgung: Bitte nur restentleerte Gebinde dem Recycling zuführen.					
GemmaX2-28-9B	Disposal: Please submit emptied containers for recycling.	69.38	**High Partial**	85.52	75
GPT-4o-mini	Disposal: Please only contribute emptied containers to recycling.	45.48	**Full match**	87.96	95
M^{2}PO	Disposal: Please only submit completely emptied containers to recycling.	76.71	**Full match**	89.52	100

Table 7: Case study: Comparison of translation alignment and quality metrics. **KIWI-XXL** and **SAC** serve as training signals, while **XCOMET** and **Coverage** are testing metrics. Blue text marks the source segments of interest. In the Translation Output, green denotes faithful translations, whereas orange and red visualize partial errors that standard QE metrics often overlook.

A Qualitative Analysis

Table 7 compares the base model (GemmaX2-28-9B), GPT-4o-mini, and M^{2}PO, illustrating how M^{2}PO leverages the **SAC** signal to rectify partial errors missed by standard rewards (KIWI-XXL).

Mitigating Hallucinations. For *En→Zh*, the base model suffers from severe code-switching ("被疏 evacuated"). Crucially, KIWI-XXL fails to penalize this, assigning a high score (80.08). In contrast, SAC flags it as "Low Partial". Guided by this, M^{2}PO eliminates the hallucination to produce a faithful translation ("被从该地区疏散"), achieving near-perfect XCOMET (99.98).

Capturing Semantic Nuances. M^{2}PO outperforms baselines in preserving semantic constraints: **De→En:** The source "nur restentleerte" requires both *only* and *completely* emptied. The base model misses both, while GPT-4o-mini overlooks "completely". M^{2}PO captures the full scope ("only submit completely emptied containers"), ensuring maximal fidelity.
Zh→En: For "口水鸡", the base model outputs misleading "cold chicken" (High Partial). M^{2}PO generates a precise translation ("Poached chicken with chili sauce"), resolving ambiguity to match the reference.

B Analysis of Metric Reliability

To validate our metric selection, we analyze reliability on the *HalOmi* benchmark (Dale et al., 2023b), tracking the performance drop (Δ) when nuanced Partial Errors are included (Table 8).

Vulnerability of Standard QE. While standard QE models align with general quality, they reveal a severe "blind spot" for partial errors. Notably, **KIWI-XXL** degrades drastically when omissions are included ($\Delta = 11.60$), and **XCOMET** follows a similar trend ($\Delta = 10.78$). This indicates a *systematic failure* of regression-based QE to penalize superficially fluent but incomplete translations, necessitating a more semantics-aware signal.

Limitations of Traditional Alignment. We explicitly compare against the continuous baseline **WSPAlign**. Despite capturing lexical correspondences, its correlation with human judgment lags significantly behind LLM-based signals (e.g., 61.97 vs. 73.99 for SAC in hallucination). This

Metric	Score Distribution (Avg. Score ↑)						Correlation Robustness (Pearson $\times 100$ ↑)					
	Hallucination			Omission			Hallucination			Omission		
	No	Part.	Full	No	Part.	Full	All	w/o Part.	Δ	All	w/o Part.	Δ
Word Alignment Baselines												
WSPAlign	64.4	45.9	25.2	69.3	51.2	25.6	61.97	64.86	+2.89	63.70	72.63	+8.94
Training Signals												
SAC	54.3	88.1	76.7	64.6	77.9	77.8	73.99	75.59	+1.60	71.22	78.54	+7.32
KIWI-XXL	63.7	24.3	9.8	64.3	36.1	9.4	61.37	64.85	**+3.48**	54.27	65.87	**+11.60**
Testing Metrics												
Coverage	85.0	48.5	9.3	81.0	60.5	8.7	**76.65**	**77.12**	+0.47	**71.74**	**79.69**	+7.95
XCOMET	76.8	36.4	16.4	79.5	52.9	16.1	69.72	72.60	+2.88	62.42	73.14	+10.78

Table 8: Analysis of metric reliability on HalOmi. We compare our training signal (SAC) against testing metrics (Coverage, XCOMET) and word alignment baselines (WSPAlign (Wu et al., 2023)). **Left:** Average quality scores (for SAC, values denote per-tag accuracy percentage). **Right:** Pearson correlation with human labels.

suggests that traditional alignment tools lack the *deep semantic reasoning* required to identify subtle semantic deviations.

Robustness of LLM-based Signals. In contrast, LLM-driven metrics demonstrate superior robustness. Specifically for hallucination evaluation, **Coverage** achieves the highest correlation (76.65) with negligible sensitivity to partial errors ($\Delta = 0.47$), validating it as our rigorous testing metric. Crucially, our proposed **SAC** (used for training) also maintains competitive reliability in detecting hallucinations (73.99), confirming that *discrete semantic diagnosis* provides a more accurate and stable supervision signal than traditional continuous baselines.

C Implementation Details: Prompts, SAC Configuration, and Costs

This section details the prompts used for data synthesis and evaluation, the configuration of the Semantic Alignment Classifier (SAC), and the cost analysis for dataset construction and evaluation.

C.1 Prompt Templates

Figure 7 presents the prompt templates used in our study. For **faithfulness assessment**, we employ the *Coverage Calculation* prompt to query hallucination/omission scores and the *SAC* prompt for discrete faithfulness labeling. For **translation**, to isolate algorithmic gains, we enforce a strict control strategy where external baselines (proprietary and open-source) follow default instructions, while M^{2}PO and controlled baselines share an identical native task format (based on GemmaX2-28-9B). This translation prompt remains invariant across sampling, training, and inference, ensuring all improvements are driven by our alignment algorithm rather than prompt engineering.

C.2 SAC Formulation and Sensitivity Analysis

To compute the alignment score S_{align} (Eq. 2), we utilize GPT-4o-mini. Beyond its cost-efficiency and instruction-following capabilities, this choice strategically decouples training supervision from evaluation, as our final faithfulness benchmarking relies on Gemini-2.0-Flash. This separation mitigates the risk of the model "gaming" a specific evaluator's biases. Specifically, the model predicts a semantic category c, which is subsequently mapped to a discrete scalar via the following schedule:

$$S_{\text{align}} = \mathcal{M}(c) = \begin{cases} 1.0 & \text{if } c \text{ is } \textit{Full Match} \\ 0.7 & \text{if } c \text{ is } \textit{High Partial} \\ 0.3 & \text{if } c \text{ is } \textit{Low Partial} \\ 0.1 & \text{if } c \text{ is } \textit{No Match} \end{cases} \tag{7}$$

Sensitivity Analysis. To validate our heuristic coefficients, we benchmark our *Stepwise* strategy against three variants: Strict Binary (zero-tolerance), Relaxed Penalty (milder sanctions), and a Zero Floor baseline. As shown in Table 9, *Strict Binary* yields high Coverage (97.90) but degrades quality (89.02 XCOMET) by discarding useful partial signals. Conversely, *Relaxed Penalty* fails to sufficiently curb hallucinations. Crucially, the *Zero Floor* setting, which strictly assigns zero weight to severe errors, underperforms our approach. This

Configuration	Weights	Coverage	XCOMET
Strict Binary	$\{1.0, 0.0, 0.0, 0.0\}$	**97.90**	89.02
Relaxed Penalty	$\{1.0, 0.9, 0.5, 0.1\}$	97.53	89.55
Zero Floor	$\{1.0, 0.7, 0.3, 0.0\}$	97.71	89.58
Ours (Stepwise)	$\{1.0, 0.7, 0.3, 0.1\}$	97.87	**89.67**

Table 9: Sensitivity analysis of SAC weight configurations on the WMT23 benchmark.

confirms that a non-zero lower bound acts as a vital *safety floor*, providing robustness against potential classifier noise while maintaining gradient flow. Ultimately, our strategy achieves the optimal balance, attaining the highest XCOMET (89.67) with near-top faithfulness (97.87).

C.3 Cost Analysis

We analyze the cost efficiency of our framework to ensure reproducibility and accessibility. As detailed in Table 10, the construction of M^{2}PO-Prefer involved substantial scale, leveraging the cost-effective GPT-4o-mini for ∼110k candidate generations and ∼160k SAC assessments. The latter required processing a significant token volume (∼57M) to handle the detailed instruction definitions needed for accurate alignment categorization.

Furthermore, our evaluation phase utilized Gemini-2.0-Flash to conduct an extensive benchmarking campaign. We computed faithfulness (Coverage) scores for approx. 15 comparative systems and ablation variants across all test sets, totaling ∼320k assessment requests. Remarkably, by strategically prioritizing these high-efficiency models, the cumulative API expenditure for the entire project remained negligible (≈**$20 USD**[10]). We leverage the ultra-low pricing of GPT-4o-mini ($0.15/$0.60 per 1M input/output tokens) for reliable construction and Gemini-2.0-Flash ($0.10/$0.40 per 1M input/output tokens) for large-scale evaluation. This demonstrates that M^{2}PO is highly accessible to the research community, enabling rigorous, large-scale alignment research without requiring prohibitive budgets.

D Dataset Statistics and Scaling Analysis

Dataset Construction. Table 11 summarizes the raw seed corpora (∼28,000 source sentences),

[10]Pricing estimates are based on official documentation as of December 2025: OpenAI (https://platform.openai.com/docs/pricing) and Google (https://ai.google.dev/gemini-api/docs/pricing).

Stage	Usage Statistics		Expenditure
	Volume	Tokens	Cost (USD)
Construction (GPT-4o-mini)			
Candidate Gen.	110k	∼16.5M	4.29
SAC Assessment†	160k	∼56.8M	8.90
Evaluation (Gemini-2.0-Flash)			
Faithfulness Eval.‡	320k	∼65.6M	7.04
Total Project	-	∼138.9M	≈ **20.23**

Table 10: Cost breakdown of the M^{2}PO framework. Expenditures are calculated based on Dec 2025 pricing for GPT-4o-mini ($0.15/$0.60 per 1M tokens for input/output) and Gemini-2.0-Flash ($0.10/$0.40). †: High token usage reflects detailed instruction prompts. ‡: Covers extensive evaluation of approx. 15 comparative systems across all benchmarks.

which are strictly isolated from testing benchmarks to ensure zero data leakage. To construct the final M^{2}PO-Prefer dataset, we expand these seeds via hybrid generation ($K = 8$) and apply rigorous curation to maximize signal clarity. Specifically, we perform (1) *deduplication* of redundant model outputs and (2) *margin filtering* based on QE scores to exclude ambiguous pairs with insufficient quality separation ($\Delta_{\text{qe}} < 2$). This pipeline yields a refined set of 20,000 high-quality instances (comprising 160,000 total candidates), prioritizing discriminative power over raw volume.

Scaling Analysis. To validate the sufficiency of our curated data scale, we investigate the relationship between training volume and performance. Figure 6 illustrates the scaling curves for Coverage and XCOMET as the proportion of M^{2}PO-Prefer increases from 0% to 100%.

The results underscore remarkable data efficiency: utilizing merely 25% of the refined set (∼5,000 samples) leads to a sharp performance leap, achieving a +1.86 gain in XCOMET and +0.99 in Coverage. This initial surge captures more than half of the total improvement observed at full scale. Beyond this point, the performance follows a logarithmic scaling pattern with diminishing marginal returns. The highly synchronized growth of both metrics confirms that the multi-pair objective effectively extracts robust alignment signals even from a limited pool of high-quality preference pairs, justifying our focus on data quality during construction.

	Dataset	Lang	$\|S\|$	$\|T\|$	$\|L\|$
Training	WMT22	En↔Zh	3.9K	82.4K	21
		En↔De	4.0K	65.9K	16
		En↔Ja	4.0K	63.1K	16
	IWSLT 15–17	En↔Zh	2.2K	35.7K	16
		En↔De	4.0K	67.7K	17
		En↔Ja	4.0K	67.2K	17
	FLORES-200-dev	En↔Zh	2.0K	41.9K	21
		En↔De	2.0K	41.9K	21
		En↔Ja	2.0K	41.9K	21
Testing	WMT23	En↔Zh	4.1K	82.8K	20
		En↔De	1.1K	60.6K	55
		En↔Ja	4.1K	68.6K	17
	FLORES-200-test	En↔Zh	2.0K	43.8K	22
		En↔De	2.0K	43.8K	22
		En↔Ja	2.0K	43.8K	22
	WMT24	En→Zh	1.0K	32.3K	32
		En→De	2.0K	64.7K	32
		En→Ja	2.0K	64.7K	32

Table 11: Statistics of the datasets employed in this study. We distinguish between *Training Sources* (seed data for preference construction) and disjoint *Testing Benchmarks*, which cover General (WMT), Spoken (IWSLT), and Encyclopedic (FLORES) domains. Reported metrics include sentence count ($|S|$), English token count ($|T|$), and average sentence length ($|L|$).

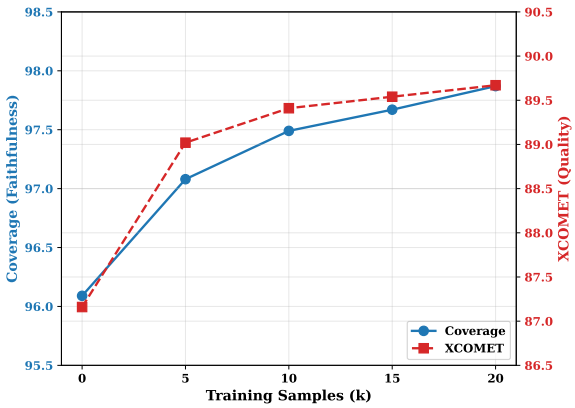

Figure 6: Scaling analysis of M^{2}PO on varying proportions of the training set.

Model / Method	Coverage	XCOMET
Base (GemmaX2-28-9B)	96.09	87.16
SFT Baselines		
+ M^{2}PO-Prefer (Top-ranked)	96.81	87.79
+ Gold Reference	97.06	87.72
Preference Optimization (Ours)		
+ M^{2}PO	**97.87**	**89.67**

Table 12: Comparison of SFT baselines trained on different targets versus M^{2}PO.

E Extended Ablation on SFT Target Construction

To determine whether our Supervised Fine-Tuning (SFT) baseline is bottlenecked by using top-ranked model candidates, we evaluate an alternative SFT model trained strictly on authentic *Gold References*. As Table 12 shows, while the Gold Reference naturally yields higher faithfulness (Coverage: 97.06 vs. 96.81 for top-ranked), M^{2}PO still consistently outperforms this optimal SFT setup. This gap highlights a fundamental limitation of Maximum Likelihood Estimation (MLE): even given perfect human references, standard SFT merely mimics the target distribution. It lacks the discriminative capability to actively penalize fine-grained partial errors (e.g., subtle omissions) that our multi-pair preference objective effectively mitigates.

F Additional Experimental Results

To rigorously verify the robustness and generalization capabilities of M^{2}PO, we extend our evaluation beyond the primary benchmark. We incorporate **FLORES-200-test** to assess domain transfer performance on diverse topics and **WMT24** to evaluate temporal robustness against newer data distributions. Additionally, we provide supplementary reference-based evaluations using **COMET-22** across all three datasets to ensure metric consistency.

As detailed in Tables 17, 18, 19, and 20, M^{2}PO consistently outperforms strong baselines across these varied settings. The COMET-22 results on WMT23 strongly corroborate the quality gains observed in the main text. Furthermore, the strong performance on FLORES-200 confirms that our faithfulness constraints transfer effectively to out-of-distribution domains, while the WMT24 results demonstrate the model's resilience to temporal shifts. These findings collectively indicate that M^{2}PO learns generalized preference patterns rather than overfitting to specific training data or evaluation metrics.

G Comparison with Online-RL Approaches

We benchmark M^{2}PO against prominent Online-RL optimizers, specifically GRPO (DeepSeek-AI, 2025) and GSPO (Zheng et al., 2025), evaluating both empirical performance and underlying algo-

System	Translation Output	MetricX-24-XXL (↓)	KIWI-XXL (↑)	SAC
Source (de→en): Entsorgung: Bitte nur restentleerte Gebinde dem Recycling zuführen.				
GemmaX2-28-9B	Disposal: Please submit emptied containers for recycling.	4.03	69.38	**High Partial**
GPT-4o-mini	Disposal: Please only contribute emptied containers to recycling.	5.25	45.48	**Full match**
M^{2}PO (Ours)	Disposal: Please only submit completely emptied containers to recycling.	3.39	76.71	**Full match**

Table 13: Qualitative comparison of continuous reward proxies (MetricX-24-XXL and KIWI-XXL) and SAC on translations containing partial errors.

Method	Reward Signal	Coverage	XCOMET	Cost
Base	N/A	96.09	87.16	-
Online-RL (Group Size = 8)				
GRPO	KIWI-XXL only	97.14	89.01	8×
GRPO	Augmented Reward	97.55	89.15	8×
GSPO	KIWI-XXL only	97.35	88.94	8×
GSPO	Augmented Reward	97.76	89.28	8×
Offline (Ours)				
M^{2}PO	Augmented Reward	**97.87**	**89.67**	**1×**

Table 14: Fair comparison between M^{2}PO and Online-RL methods. All methods in the comparison use the same composite reward signal to isolate the impact of the optimization framework. *Cost* denotes relative training time.

rithmic mechanisms.

Empirical Superiority: Quality and Efficiency. To isolate algorithmic gains from reward quality, we align the reward proxies across all baselines. As shown in Table 14, while the **Augmented Reward** (KIWI-XXL+SAC+Model Confidence) enhances all models, M^{2}PO consistently maintains a superior performance frontier. Under identical reward conditions, it yields improvements ranging from 0.11 to 0.32 in Coverage and 0.39 to 0.52 in XCOMET. Beyond quality gains, M^{2}PO bypasses the online sampling bottleneck and the complexity of group-wide advantage calculations, achieving an 8× speedup in GPU hours compared to Online-RL baselines. This dual advantage in both generation fidelity and computational cost establishes M^{2}PO as a highly scalable and effective alignment paradigm.

Mechanistic Analysis: Structured Contrast vs. Group Mean. The performance gap between M^{2}PO and group-based counterparts highlights a fundamental difference in how these algorithms utilize candidate sets. Group-based methods (GRPO/GSPO) treat the candidate pool as an unstructured set, optimizing each candidate relative to a uniform group average ($r_i - \bar{r}$). While effective for general alignment, this aggregation inherently dilutes the specific, fine-grained contrast between a high-quality translation and a targeted "hard negative" (e.g., a fluent but partially unfaithful candidate). In contrast, M^{2}PO is explicitly designed to exploit the ordered, hierarchical structure of the candidate list. Rather than evaluating against an aggregated baseline, M^{2}PO's pairwise coupling isolates the exact semantic boundaries the model needs to learn. By structurally enforcing a direct contrast between specific preferred and rejected pairs, it extracts a more precise and targeted learning signal for subtle translation errors than a generalized group mean can provide.

H Qualitative Analysis of Metric Reliability

To address whether advanced regression metrics like MetricX-24-XXL can eliminate the need for explicit faithfulness supervision, Table 13 illustrates their limitations when evaluating partial errors.

The Blind Spot of Continuous Metrics. Continuous metrics often exhibit a critical blind spot: prioritizing surface-level fluency over strict semantic faithfulness. In the *de→en* case, GemmaX2-28-9B commits a *Major Omission* (missing "nur" and "restentleerte"). Yet, it receives a "better" distance score from MetricX-24-XXL (4.03) than GPT-4o-mini (5.25), which only has a *Minor Omission*. KIWI-XXL shows a similar ranking inversion (69.38 vs. 45.48). This confirms that even leading metrics like MetricX-24-XXL fail to reliably penalize unfaithful translations if the output remains highly fluent.

λ_{rank} Value	Coverage	XCOMET
0.0 *(w/o $\mathcal{L}_{Rank}$)*	97.15	89.19
0.1	97.58	89.45
0.5 (Ours)	**97.87**	**89.67**
1.0	97.63	89.51
2.0	97.28	89.24

Table 15: Sensitivity analysis of the ranking loss weight λ_{rank} on the WMT23 benchmark.

The Necessity of SAC. Relying solely on such metrics provides misleading training signals that favor fluent omissions. Our SAC classifier acts as an essential semantic guardrail. As shown, it accurately detects the severe semantic loss in the GemmaX2-28-9B output (assigning a High Partial penalty) while correctly recognizing GPT-4o-mini as a Full match. By integrating this discrete supervision, our framework successfully overrides the ranking biases of continuous metrics, ensuring optimization is strictly anchored to semantic faithfulness.

I Impact of Ranking Regularization Weight (λ_{rank})

We conduct a sensitivity analysis on the ranking regularization weight λ_{rank} to determine its optimal contribution to the joint objective. Table 15 summarizes the performance on WMT23 across a value spectrum of $\{0.0, 0.1, 0.5, 1.0, 2.0\}$.

The results exhibit a clear inverted U-shaped trend. Starting from the baseline without ranking regularization ($\lambda_{\text{rank}} = 0.0$), we observe steady improvements as the weight increases. Specifically, the configuration with $\lambda_{\text{rank}} = 0.5$ reaches the performance peak, achieving significant gains of +0.72 in Coverage and +0.48 in XCOMET compared to the baseline. This confirms that incorporating global ordinal consistency effectively complements the local pairwise optimization, helping the model distinguish fine-grained quality differences.

However, the performance begins to degrade when the weight exceeds this threshold ($\lambda_{\text{rank}} \geq 1.0$). This decline suggests that an overly aggressive ranking penalty creates a soft constraint that competes with the primary CPO objective. When the ranking loss dominates, it forces the model to rigidly fit the target distribution rather than focusing on maximizing the decision margin for the optimal translation. Consequently, we adopt $\lambda_{\text{rank}} = 0.5$ as the default setting to balance global consistency with local discriminative power.

Configuration	Performance	
	Coverage	XCOMET
Base (GemmaX2-28-9B)	96.09	87.16
Fixed Strategies (Static Weight)		
$\alpha \equiv 0.0$ (Static Only)	96.98	89.44
$\alpha \equiv 0.5$ (Equal Mix)	97.01	89.08
$\alpha \equiv 1.0$ (Conf. Only)	96.67	88.32
Dynamic Schedules (α_t : Start $\rightarrow$ End)		
$0.3 \rightarrow 0.7$ (Narrow)	97.35	89.48
$0.0 \rightarrow 1.0$ (Full Range)	97.62	89.55
$0.1 \rightarrow 0.9$ **(Clipped)**	**97.87**	**89.67**

Table 16: Ablation study on the curriculum schedule α_t on the WMT23. We compare fixed scalar weights against dynamic schedules with varying boundaries.

J Impact of Dynamic Weighting Strategy (α_t)

To investigate the sensitivity of our dynamic curriculum (α_t) to boundary constraints, we compare the proposed schedule against fixed and alternative dynamic ranges. Table 16 confirms that while dynamic scheduling generally outperforms fixed baselines, the choice of boundary values is pivotal. The *Narrow Schedule* ($0.3 \rightarrow 0.7$) underperforms due to insufficient decoupling of external anchoring and internal refinement. While the *Full Range* ($0.0 \rightarrow 1.0$) enhances quality, it suffers from faithfulness regression driven by "self-delusion"—unchecked hallucinations at $\alpha_t = 1.0$. Consequently, our *Clipped Schedule* ($0.1 \rightarrow 0.9$) strikes the optimal balance; retaining a 10% *safety anchor* ensures semantic fidelity while maximizing self-refinement.

Prompt of Coverage Calculation

Given a source sentence {source_text} in {source_language} and a translation {translation} in {target_language}, does the translation has hallucination or omission to the source contents?

You MUST answer with only a coverage percentage score.

Prompt of Semantic Alignment Categorization

Classify the translation alignment into one category:
- [Full match]: Semantic and constituent equivalent.
- [High Partial match]: Core meaning preserved, but contains minor errors or omissions.
- [Low Partial match]: Core meaning lost or significantly changed, but contains valid fragments/keywords.
- [No match]: Completely unrelated or hallucinated.

Input:
Direction: {source_language} to {target_language}
source_text: {source_text}
translation: {translation}

Output ONLY the label.

Translation Prompt of External Baselines

System:
You are a helpful translator and only output the result.

User:
Translate this from {source_language} to {target_language}:
{source_language}: {source_text}
{target_language}:

Translation Prompt of Controlled Baselines & M^{2}PO

Translate this from {source_language} to {target_language}:
{source_language} : {source_text}
{target_language}:

Figure 7: **Overview of prompt templates. Top:** Prompts for faithfulness assessment. The Coverage Calculation serves as the primary metric for *test evaluation*, while the Semantic Alignment Categorization (SAC) is employed for constructing preference pairs. **Bottom:** The inference prompts for translation tasks, distinguishing between the format for *External Baselines* and the native format for our *Base Model*, *Controlled Baselines*, and *M^{2}PO*.

Model	AVG	En→Zh	En→De	En→Ja	Zh→En	De→En	Ja→En
Gemini-2.0-Flash	84.82	86.20	84.50	**88.11**	80.87	85.86	**83.37**
GPT-4o	**84.92**	86.57	**84.53**	87.68	81.33	**86.17**	83.23
GPT-4o-mini	84.66	86.33	84.15	87.48	81.31	85.70	82.96
Aya-expanse-32B	84.11	86.31	82.10	87.70	80.62	85.22	82.69
Aya-23-35B	82.81	84.74	80.66	86.44	79.37	84.15	81.49
TowerInstruct-13B	80.84	84.98	79.28	77.08	80.16	84.38	79.17
ALMA-13B-R	81.98	83.99	79.88	83.77	80.42	84.57	79.27
Qwen3-4B-Instruct	81.42	84.10	76.88	84.27	80.41	82.57	80.27
+ SFT	82.24	85.28	77.13	84.96	80.88	84.08	81.13
+ CPO	82.78	86.39	77.47	86.05	81.00	84.11	81.66
+ M^{2}PO	83.07	86.47	78.19	86.37	81.16	84.35	81.87
GemmaX2-28-9B	83.36	85.99	81.59	85.91	80.52	84.73	81.43
+ SFT	83.74	86.64	81.77	86.31	80.87	84.94	81.93
+ CPO	84.03	86.73	81.98	86.82	81.18	85.11	82.37
+ M^{2}PO	84.61	**86.93**	82.34	88.06	**81.63**	85.42	83.26

Table 17: Main results on the WMT23 benchmark. Results are reported in **COMET22** scores. Model names are color-coded by type: Proprietary, Open-source baselines, and Our experiments. **Bold** indicates the best result overall. Colored background indicates the best result among open-source models.

Model	AVG	En→Zh	En→De	En→Ja	Zh→En	De→En	Ja→En
Gemini-2.0-Flash	98.04 / 95.10	97.29 / 91.97	99.05 / 98.01	97.74 / 93.29	98.20 / 96.49	98.79 / 96.27	97.17 / 94.56
GPT-4o	**98.42** / 95.21	97.62 / 91.20	**99.43** / 98.18	98.39 / 93.43	98.58 / 96.73	**98.96** / 96.64	**97.55** / **95.06**
GPT-4o-mini	98.30 / 94.66	97.38 / 90.01	99.37 / 97.74	98.26 / 92.95	98.56 / 96.54	98.84 / 96.22	97.36 / 94.50
Aya-expanse-32B	97.92 / 95.02	97.05 / 91.87	99.11 / **98.39**	97.73 / 93.63	98.11 / 96.64	98.55 / 96.10	96.98 / 93.48
Aya-23-35B	97.46 / 94.10	96.67 / 90.90	98.87 / 98.15	97.06 / 92.30	97.59 / 96.16	98.28 / 94.99	96.30 / 92.08
TowerInstruct-13B	97.25 / 93.03	96.45 / 89.60	99.12 / 98.05	95.87 / 86.16	97.54 / 96.14	98.32 / 95.94	96.19 / 92.39
ALMA-13B-R	96.39 / 92.20	94.77 / 89.32	98.22 / 98.21	94.60 / 81.88	97.25 / 96.25	98.49 / 96.57	94.98 / 90.99
Qwen3-4B-Instruct	96.13 / 92.66	96.14 / 89.63	95.88 / 96.65	96.82 / 90.53	97.09 / 94.81	96.68 / 94.14	94.14 / 90.17
+ SFT	96.64 / 93.86	96.53 / 90.91	97.15 / 97.45	96.87 / 91.64	97.15 / 95.34	97.44 / 95.68	94.68 / 92.11
+ CPO	97.15 / 94.14	97.01 / 91.28	97.05 / 97.82	97.26 / 91.93	97.46 / 95.82	97.82 / 95.52	96.32 / 92.49
+ M^{2}PO	98.03 / 94.76	97.32 / 91.84	98.95 / 98.10	97.98 / 92.49	98.21 / 96.39	98.62 / 95.96	97.12 / 93.77
GemmaX2-28-9B	96.87 / 93.63	96.91 / 90.69	98.07 / 97.67	96.60 / 90.95	97.30 / 95.47	97.23 / 95.28	95.13 / 91.69
+ SFT	97.66 / 94.68	97.20 / 90.69	98.71 / 97.53	97.26 / 92.85	98.15 / 96.64	98.12 / 96.37	96.53 / 93.99
+ CPO	97.76 / 95.18	97.12 / 92.13	98.75 / 97.24	97.38 / 93.60	98.21 / 96.73	98.25 / 96.65	96.83 / 94.70
+ M^{2}PO	98.41 / **95.73**	**97.68** / **92.51**	99.37 / 98.23	**98.40** / **94.50**	**98.76** / **97.16**	98.93 / **96.98**	97.34 / 94.98

Table 18: Main results on the FLORES-200-test benchmark. Results are reported in the format of **Coverage / XCOMET**. Model names are color-coded by type: Proprietary, Open-source baselines, and Our experiments. **Bold** indicates the best result overall. Colored background indicates the best result among open-source models.

Model	AVG	En→Zh	En→De	En→Ja	Zh→En	De→En	Ja→En
Gemini-2.0-Flash	89.30	89.04	88.93	91.65	87.71	89.79	88.67
GPT-4o	89.32	88.91	**88.98**	91.76	87.78	**89.92**	88.55
GPT-4o-mini	88.88	88.31	88.42	91.36	87.32	89.69	88.20
Aya-expanse-32B	89.22	88.85	88.48	91.64	**87.80**	89.83	**88.73**
Aya-23-35B	88.60	87.73	87.91	90.96	87.44	89.42	88.11
TowerInstruct-13B	88.40	88.40	88.23	89.04	87.33	89.64	87.75
ALMA-13B-R	87.80	87.07	87.96	88.37	87.04	89.50	86.84
Qwen3-4B-Instruct	87.40	86.73	87.08	88.77	86.72	88.15	86.94
+ SFT	88.06	87.38	87.00	90.13	87.25	89.08	87.53
+ CPO	88.52	87.67	87.90	90.23	87.28	89.85	88.17
+ M²PO	88.88	88.91	88.29	90.86	87.37	89.44	88.38
GemmaX2-28-9B	88.48	88.60	87.54	90.50	86.93	89.20	88.10
+ SFT	88.75	88.38	87.90	91.23	87.28	89.55	88.17
+ CPO	88.87	88.61	87.72	91.50	87.39	89.68	88.29
+ M²PO	**89.33**	**89.11**	88.86	**91.93**	87.67	89.79	88.61

Table 19: Main results on the FLORES-200-test benchmark. Results are reported in **COMET22** scores. Model names are color-coded by type: Proprietary, Open-source baselines, and Our experiments. **Bold** indicates the best result overall. Colored background indicates the best result among open-source models.

Model	AVG	En→Zh	En→De	En→Ja
Gemini-2.0-Flash	97.11 / 86.38 / 84.28	96.42 / 81.80 / 83.86	98.47 / 92.56 / 82.52	96.45 / 84.79 / 86.47
GPT-4o	97.25 / 86.00 / 84.03	96.57 / 80.91 / 83.66	**98.56** / 92.83 / 82.58	96.63 / 84.26 / 85.84
GPT-4o-mini	97.15 / 84.95 / 83.88	96.63 / 79.96 / 83.15	98.24 / 90.83 / 82.16	96.59 / 84.07 / 86.33
Aya-expanse-32B	96.81 / 86.45 / 84.42	96.44 / 81.21 / 83.56	97.90 / 92.74 / 82.56	96.09 / 85.40 / **87.14**
Aya-23-35B	95.95 / 84.36 / 82.87	95.20 / 79.66 / 81.84	97.79 / 91.78 / 81.36	94.87 / 81.64 / 85.42
TowerInstruct-13B	94.49 / 76.45 / 80.69	94.33 / 74.50 / 81.16	97.81 / 91.23 / 81.44	91.33 / 63.63 / 79.47
ALMA-13B-R	92.68 / 74.84 / 80.30	92.93 / 73.38 / 80.53	96.77 / 91.46 / 81.25	88.35 / 59.68 / 79.13
Qwen3-4B-Instruct	95.61 / 82.31 / 81.74	96.09 / 80.60 / 82.98	95.76 / 88.29 / 78.41	94.97 / 78.04 / 83.82
+ SFT	95.96 / 82.87 / 82.25	96.44 / 81.45 / 83.38	96.30 / 88.72 / 79.25	95.15 / 78.43 / 84.12
+ CPO	96.51 / 83.20 / 82.63	96.57 / 81.84 / 83.98	97.13 / 89.02 / 79.51	95.84 / 78.75 / 84.39
+ M²PO	96.95 / 83.76 / 83.11	96.78 / 82.54 / 84.36	97.72 / 89.39 / 80.10	96.36 / 79.36 / 84.86
GemmaX2-28-9B	95.65 / 84.69 / 83.81	94.79 / 80.56 / 84.11	96.94 / 92.19 / 82.28	95.23 / 81.32 / 85.03
+ SFT	96.54 / 86.28 / 84.32	96.13 / 81.52 / 84.15	97.40 / 92.06 / 82.32	96.09 / 85.27 / 86.49
+ CPO	96.72 / 86.71 / 84.51	96.29 / 82.09 / 84.35	97.59 / 92.31 / 82.42	96.28 / 85.73 / 86.77
+ M²PO	**97.31** / **87.34** / **84.92**	**96.83** / **82.94** / **84.87**	98.32 / **92.84** / **82.89**	**96.77** / **86.24** / 87.00

Table 20: Main results on the WMT24 benchmark (En→X). Results are reported in the format of **Coverage / XCOMET / COMET22**. Model names are color-coded by type: Proprietary, Open-source baselines, and Our experiments. **Bold** indicates the best result overall. Colored background indicates the best result among open-source models.