Remedy-R: Generative Reasoning for Machine Translation Evaluation without Error Annotations

Shaomu Tan[1,2,*] **Ryosuke Mitani**[2] **Ritvik Choudhary**[2] **Qiyu Wu**[2]
Toshiyuki Sekiya[2] **Christof Monz**[1]

[1]University of Amsterdam [2]Sony Group Corporation
s.tan@uva.nl

Abstract

Over the years, scalar MT metrics have advanced rapidly on benchmarks. Yet they remain black boxes, offering little insight into their decisions and sometimes degrading under out-of-distribution inputs. We introduce Remedy-R, a reasoning-driven generative MT metric trained with reinforcement learning from pairwise translation preferences, without requiring error-span annotations or distillation from closed LLMs. Unlike scalar MT metrics that only output translation quality scores, Remedy-R produces step-by-step analyses of accuracy, fluency, and completeness, enabling more interpretable assessments. With only 60K pairwise training samples across two language pairs, Remedy-R remains competitive with top scalar metrics and GPT-4-based judges on WMT22–25 metric benchmarks, generalizes to other languages, and shows strong robustness on OOD stress tests. Moreover, Remedy-R generates self-reflective feedback that can be reused for translation refinement. We validate the faithfulness of such feedback with GPT-4 and show that a simple evaluate–revise pipeline leveraging Remedy-R's analyses consistently improves translation quality across diverse models without any task-specific tuning.[1]

1 Introduction

Recent neural machine translation (MT) evaluation metrics like xCOMET (Guerreiro et al., 2024), MetricX (Juraska et al., 2024), and Remedy (Tan and Monz, 2025) achieve strong correlations with human preferences, and in some settings even report agreement that exceeds expert annotators (Proietti et al., 2025). Yet, they remain black boxes, producing a single scalar score with little insight into why a translation is good or bad.

This opacity matters because translation quality is multi-dimensional, involving criteria like accuracy, fluency, and completeness. A single number does not reveal which dimension drives the judgment, and it also complicates robustness: without explicit decision-making process, metrics may exploit spurious cues learned during training and degrade under out-of-distribution (OOD) inputs such as source copy and input perturbations (Lo et al., 2023; Knowles et al., 2024; Moghe et al., 2025). As a result, existing MT metrics are powerful but hard to interpret and diagnose.

Recent work has attempted to enhance interpretability through translation error-span prediction, highlighting words or phrases that contribute to errors (Guerreiro et al., 2024; Treviso et al., 2024). Despite low overlap with human-annotated spans, such predictions are inherently local. In particular, span-based feedback struggles with omissions, discourse-level issues, and sentence-level phenomena not well captured by span annotations (e.g., translations in wrong or mixed languages): a translation may have few or no highlighted spans yet still be inadequate overall.

Another direction uses LLMs-as-judges to perform multi-dimensional assessment like MQM (Freitag et al., 2021). However, many LLMs-as-judges approaches still reduce the analysis to a final quality score (Kocmi and Federmann, 2023b; Fernandes et al., 2023; Qian et al., 2024), and existing works often rely on closed commercial LLMs (Lu et al., 2024; Kocmi and Federmann, 2023a; Zhang et al., 2025; Feng et al., 2025b), while open models struggle to match them (Lu et al., 2025). This raises a key question: how can we train an open, reasoning-based MT metric that produces informative analyses and aligns with human preferences, without expensive error annotations?

We introduce *Remedy-R*, a generative MT metric that produces step-by-step analyses along three dimensions (accuracy, fluency, completeness) and a final score. Inspired by recent pairwise preference framework in Remedy (Tan and Monz,

[*] Work done while interning at Sony.
[1] https://github.com/Smu-Tan/Remedy-R

Findings of the Association for Computational Linguistics: ACL 2026, pages 7374–7398

Remedy-R ref-free Evaluation Example

Source: Even if I'm not in the income level that would pay more taxes, I can still buy fewer things that are more expensive so that they can be made by people making a living wage and I can be willing to wait a few days to get it so that some gig worker doesn't have to be worked to the bone...

Translation: Sogar wenn ich nicht in der Einkommensklasse bin, in der mehr Steuern gezahlt werden, kann ich immer noch weniger teure Dinge kaufen, sodass sie von Menschen hergestellt werden können, die maßgeblich dazu beitragen, und ich bin bereit, ein paar Tage zu warten, bevor der Arbeitslohn abläuft...

Remedy-R's Evaluation:
Let's evaluate the translation step by step:

- **Accuracy**
  - "sodass sie von Menschen hergestellt werden können, die maßgeblich dazu beitragen" is not clear and does not accurately reflect the original meaning. The original text states that the products should be made by people who are paid a living wage, but the translation does not convey this clearly.
  - "bevor der Arbeitslohn abläuft" is incorrect and does not make sense in this context. The original text states "so that some gig worker doesn't have to be worked to the bone" which means the gig worker is not overworked. The translation should reflect this meaning.
- **Fluency**
  - The sentence structure is somewhat complex and could be clearer.
  - There are some awkward phrasings, such as "in der Einkommensklasse bin, in der mehr Steuern gezahlt werden," which could be simplified.
  - The use of "maßgeblich dazu beitragen" is not clear and does not fit the context.
- **Completeness**
  - The original text also states that the speaker is willing to wait a few days to get the product so that a gig worker doesn't have to be overworked. This part is also not clearly conveyed in the translation.

Score: 65.

2025), Remedy-R is trained with Reinforcement Learning via Verifiable Rewards (RLVR) from *pairwise human preferences*. Specifically, we define a deterministic reward that checks whether the model-implied preference (from parsed final scores) matches the human preference label, and optimize Remedy-R with PPO (Schulman et al., 2017) to maximize verifiable rewards. Unlike xTower, which depends on GPT-4 distillation and supervised learning with error spans (Treviso et al., 2024), Remedy-R trains directly with RLVR on MT preference data, without requiring error-span annotations, distillation, or SFT cold-start.

We assess whether Remedy-R's analyses are faithful and practically useful in two complementary ways. First, we validate explanation faithfulness with GPT-4o-mini (§5.1). Second, we introduce *Remedy-R Agent*, a simple training-free evaluate–revise loop that reuses Remedy-R's analyses as feedback to refine translations (§5). Despite never being trained for refinement, the agent consistently improves translation quality across diverse models (Qwen2.5 (Yang et al., 2024), ALMA-R (Xu et al., 2024b), GPT-4o-mini (Achiam et al., 2023), and Gemini-2.0-Flash) and generalizes to 11 language pairs beyond its two training language pairs. Our contributions are:

- We propose *Remedy-R*, a generative MT evaluator that produces step-by-step analyses and a final score, trained directly via RLVR without error span annotation or distillation.
- Remedy-R achieves competitive performance on WMT22–25 meta-evaluation with only 60K training pairs from two language pairs, and exhibits strong OOD behavior.
- Remedy-R Agent, as a training-free framework, improves translations across diverse model families, and maintains cross-lingual generalization beyond its training languages.

2 Method

2.1 Task Formulation

We revisit MT evaluation as follows. Given a source sentence src, a translation mt, and an optional reference ref^* (with $ref^* = \emptyset$ in reference-free settings), a metric M outputs a scalar quality score. Conventional learned metrics directly map (src, mt, ref^*) to a single number, which limits its interpretability and reusability for downstream refinement. In Remedy-R, we formulate MT evaluation as *conditional text generation* guided by

an instruction that specifies the task and criteria, namely accuracy, fluency, and completeness. The model follows a reason then score protocol: it first writes a short analysis and then outputs a numeric score that we can parse for evaluation. Here, let the input be:

$$\mathbf{x} = \langle instruct,\ src,\ mt,\ ref^{*} \rangle.$$

where the instruction *instruct* specifies the evaluation task and criteria (e.g., accuracy, fluency, completeness). The model produces an output sequence $\mathbf{y}$ in a *reason-then-score* format: a step-by-step reasoning analysis and a final numeric score in $[0, 100]$. We parameterize a conditional policy $\pi_\theta(\mathbf{y} \mid \mathbf{x})$ and generate autoregressively:

$$\pi_\theta(\mathbf{y} \mid \mathbf{x}) = \prod_{t=1}^{T} \pi_\theta\big(y_t \mid \mathbf{x}, y_{<t}\big). \tag{1}$$

Here T denotes the output length. The final score in the output is directly parsed for evaluation and will support the verifiable rewards used in §2.2.

2.2 Remedy-R Reward Design

We optimize Remedy-R with reinforcement learning using a simple, fully verifiable reward aligned with human judgments. The reward combines (i) a sparse *pairwise ranking* signal that matches the model-implied preference to the human preference label, and (ii) a reward shaping term that encourages predicted scores to be close to human scores, turning the sparse signal into a richer, continuous one. We maximize the expected reward with PPO; training details are in §2.3 (and Appendix A.1).

2.2.1 Pairwise ranking reward

For each source sentence, we have two translations mt_A, mt_B with human scores $g_A, g_B \in [0, 100]$. We define the human preference label by comparing g_A and g_B and exclude ties ($g_A = g_B$). We instruct the model to produce a reasoning COTs path and two final quality scores $s_A, s_B \in [0, 100]$ in a fixed format (see the training template below). Instead of predicting a ranking label directly, we ask the model to evaluate A and B independently and to assign scores to each; this enables single segment evaluation at inference time without quadratic pairwise comparisons. We randomize the order of A/B during training to reduce position bias (Sproat et al., 2025).

The pairwise ranking reward then checks whether the model's predicted ranking matches the human ranking:

$$r_{\text{rank}} = \begin{cases} 1, & \text{if model and human rankings agree,} \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

This ranking reward is sparse and binary, but fully verifiable because it depends only on parsed rankings and human labels. If the score block cannot be parsed or falls outside the valid range, the reward is zero, and ties are treated as zero reward as well.

2.2.2 Reward shaping

The pairwise ranking reward encourages correct relative preferences, but it does not calibrate score magnitudes: many score pairs can yield the same ranking and thus the same r_{rank}. For instance, if the model outputs $(s_A, s_B) = (100, 99)$ and the human scores are $(g_A, g_B) = (100, 0)$, both cases receive $r_{\text{rank}} = 1$. This is undesirable because we ultimately want Remedy-R's numeric scores to be meaningful and comparable across translations and systems, rather than only producing correct orderings. To provide a denser learning signal while remaining robust to noisy human scores, we add a calibration term that penalizes the deviation between predicted scores and human scores.

Specifically, let $e_A = s_A - g_A$ and $e_B = s_B - g_B$. We adopt a Huber-style penalty:

$$\rho_c(e) = \begin{cases} \frac{1}{2}\frac{e^2}{c}, & |e| \le c, \\ |e| - \frac{1}{2}c, & |e| > c, \end{cases} \tag{3}$$

where c defines a tolerance region (we set c=5 for all experiments). The Huber form penalizes small errors smoothly while being less sensitive to large outliers, which is helpful given annotator noise and near-tie cases (Freitag et al., 2021; Tan and Monz, 2025). We normalize and average the penalties to obtain

$$\psi = \frac{1}{2}\left(\frac{\rho_c(e_A)}{c} + \frac{\rho_c(e_B)}{c}\right). \tag{4}$$

The final shaped reward is

$$r = r_{\text{rank}} \cdot \big(1 - \beta\,\psi\big), \tag{5}$$

where β controls shaping strength. We apply shaping only when the ranking is correct (through multiplication by r_{rank}), preserving strict verifiability while improving score calibration. Appendix A.2 presents an ablation study on reward shaping. We also tested an auxiliary explanation-quality penalty using genRM, which provided only marginal gains (see A.2). Accordingly, we adopt the ranking reward with Huber shaping for all main results.

2.3 RLVR Training with PPO

We train Remedy-R with RLVR using the verifiable reward defined in §2.2. Given an input translation pair, the model generates analyses and final scores in a fixed format; we parse the score block and compute a terminal scalar reward based on (i) agreement with the human pairwise preference and (ii) Huber-based score calibration.

We optimize the policy with Proximal Policy Optimization (PPO) (Schulman et al., 2017), using a KL penalty to a frozen reference model to stabilize updates and avoid reward over-optimization. Since rewards are terminal, we estimate token-level advantages with Generalized Advantage Estimation (GAE) and set $\lambda = 1$ following recent practice (Hu et al., 2025). Full objectives and training details are provided in Appendix A.1.

3 Experimental Setup

This section describes our training data, evaluation benchmarks, baselines, and implementation details.

3.1 Training Data

We train Remedy-R on WMT20 MQM (Mathur et al., 2020), constructing pairwise preferences from scalar scores without error-span annotations. MQM is expert-annotated and typically more reliable than crowd-sourced Direct Assessment (DA) (Freitag et al., 2021). In a preliminary experiment, we augmented training with an additional 120K DA preference pairs but observed no clear gains, suggesting that preference quality matters more than quantity. After constructing pairwise preferences from human scores, we obtain around 60K training pairs covering English–German and Chinese–English.

3.2 Meta Evaluation Benchmarks

We evaluate Remedy-R on (i) official WMT22–25 MQM metric benchmarks (Freitag et al., 2022, 2023, 2024; Lavie et al., 2025) and (ii) the MSLC-24 robustness suite (Knowles et al., 2024). WMT22–25 provides standardized MQM-based meta-evaluation covering a large number of MT systems across multiple language pairs. MSLC-24 contains controlled OOD and adversarial conditions, including empty inputs, source copy, wrong or mixed language outputs, unrelated translations, and perturbations, testing whether metrics penalize degenerate or inconsistent outputs.

3.3 Baselines

We compare Remedy-R against two main classes of strong baselines that represent common systems in MT evaluation: (i) *LLM-as-Judge* evaluators that use prompting to produce direct scores or MQM error spans, and (ii) *Scalar MT Metrics* trained to output only scalar translation quality scores.

LLM-as-Judge. We include GEMBA (Kocmi and Federmann, 2023b,a) as a widely used GPT-4-based family covering direct assessment scoring and MQM error spans, and PaLM-based judging (Fernandes et al., 2023; Chowdhery et al., 2023) as an alternative closed LLM evaluator. For open models, we compare to EAPrompt (Lu et al., 2024), which combines chain-of-thought reasoning with span-based feedback, and MQM-APE (Lu et al., 2025), which augments MQM error analysis with a filtering stage to reduce non-impactful errors.

Scalar MT metrics. We compare against COMET-22 (Rei et al., 2022), a widely adopted XLM-R-based regression metric, and MetricX-XXL (Juraska et al., 2023, 2024), a state-of-the-art mT5-XXL-based regression metric with strong performance on WMT22–24. We also include PaLM-2 BISON fine-tuning (Fernandes et al., 2023) as a large-model regression baseline, and Remedy (Tan and Monz, 2025) as the recent SOTA metric.

3.4 Implementation and Meta-Evaluation

We train Remedy-R on Qwen2.5 instruct models (7B, 14B, 32B) using VeRL (Sheng et al., 2024) with PPO and vLLM (Kwon et al., 2023) for rollouts. We use max sequence length 2048, Adam with learning rate 5×10^{-6}, and effective batch size 2048. We also experimented with *Qwen2.5-base* checkpoints and observed comparable performance, while instruct versions converged faster. We train for 2 epochs since gains plateau afterward (Figure 3), taking 10–27 hours on 4–8 H100/H200.

For meta-evaluation, we use the official WMT toolkit (MTME)[2] and report system-level pairwise accuracy (Acc) (Kocmi et al., 2021) and segment-level tie-calibrated accuracy (acc^*_{eq}) (Deutsch et al., 2023), with Perm-Both significance testing (Deutsch et al., 2021). For WMT24 and WMT25, we report Soft Pairwise Accuracy (SPA) (Thompson et al., 2024; Freitag et al., 2024).

[2] https://github.com/google-research/mt-metrics-eval

Type	Methods	θ	System-Level	Segment-Level acc^*_{eq}				Avg
			Acc (3 LPs)	Avg	En-De	En-Ru	Zh-En	Corr
Scalar Metrics	COMET-22-DA	0.5B	82.8	54.5	58.2	49.5	55.7	68.7
	COMET-22 (ensemble)	5x0.5B	83.9	57.3	60.2	54.1	57.7	70.6
	MetricX-XXL	13B	85.0	58.8	61.1	54.6	60.6	71.9
	PaLM-2 BISON FT	>100B	88.0	57.3	61.0	51.5	59.5	72.7
	Remedy	9B	91.2	58.9	61.0	60.4	55.4	75.1
LLM Judges	EAPrompt (Llama2)	70B	85.4	52.3	55.2	51.4	50.2	68.9
	EAPrompt (Mistral)	8x7B	84.0	50.9	53.8	50.6	48.2	67.5
	EAPrompt (GPT3.5-Turbo)	>100B	91.2	53.3	56.7	53.3	50.0	72.3
	GEMBA-MQM (Qwen)	72B	84.7	53.8	56.0	54.7	50.6	69.3
	MQM-APE (Qwen)	72B	85.8	54.5	56.4	55.7	51.4	70.2
	MQM-APE (Mistral)	8x22B	88.3	54.2	56.9	55.1	50.6	71.3
	GEMBA-DA (GPT4)	>100B	89.8	55.6	58.2	55.0	53.4	72.7
	PaLM	540B	90.1	50.8	55.4	48.6	48.5	70.5
	Remedy-R	7B	89.1	54.8	58.0	56.0	50.4	71.9
	Remedy-R	14B	88.7	56.0	58.0	55.8	54.2	72.4
	Remedy-R	32B	91.6	55.2	57.8	55.7	52.2	73.4

Table 1: Evaluation on WMT22 MQM set. Following official WMT22 settings, we report system-level Pairwise Accuracy (Acc) and segment-level pairwise accuracy with tie calibration (acc^*_{eq}), using Perm-Both statistical significance test (Deutsch et al., 2021). orange and blue indicate the best performing scalar and LLM judge metrics. Our Remedy-R 7B outperforms all LLM judges with 70B parameters, and Remedy-R 32B surpasses GEMBA-DA (GPT4).

4 Automatic Translation Evaluation Results and Analyses

In this section, we evaluate Remedy-R on automatic translation evaluation tasks and address three questions: (a) how Remedy-R compares with top scalar MT metrics and LLM-as-Judge (§4.1); (b) how performance changes under test-time-scaling with multiple evaluation passes (§4.2); and (c) whether the model behaves robustly on out-of-domain stress tests (§4.3).

4.1 Meta Evaluation on WMT benchmarks

WMT22. As shown in Table 1, Remedy-R achieves leading performance on the WMT22 metric benchmark compared to both commercial LLM judges and open metrics. Specifically, Remedy-R 7B surpasses all open LLM-as-Judge baselines with 70B parameters (e.g., EAPrompt and MQM-APE), while Remedy-R 32B exceeds GEMBA-DA (GPT-4). Among scalar metrics, Remedy-R also outperforms the regression-based PaLM-2 BISON and MetricX-XXL, showing that our model aligns with human ratings on human translation quality judgments.

WMT23-25. In addition, we also provide additional results on WMT23, WMT24, and WMT25 metric benchmarks in Table 11, Table 7, and Figure 1. The results highlight that Remedy-R achieves on-par or superior performance to the current SOTA metrics. For example, on WMT23, Remedy-R-7B outperforms EAPrompt that is based on GPT-4o-mini, and Remedy-R 14B and 32B performs on par with KIWI-XXL and MetricX-23. On WMT24, Remedy-R-14B outperforms GEMBA-ESA (GPT4), MetricX-24-Hybrid, and XCOMET-XXL (see details in Table 7). On the newest WMT25, Remedy-R 32B, equipped with TTS (8 turns) performs on-par with OpenAI-o3, while outperforming GPT-4.1-mini (10 turns).

4.2 Test Time Scaling with multiple Evaluation Passes

Remedy-R's generative reasoning nature enables the application of *Test-Time Scaling (TTS)*, where multiple evaluation passes are performed with different reasoning trajectories and their quality scores are aggregated. In this setting, we adopt a simple implementation that averages the quality scores from multiple independent evaluations.

We found performing more evaluation trajecto-

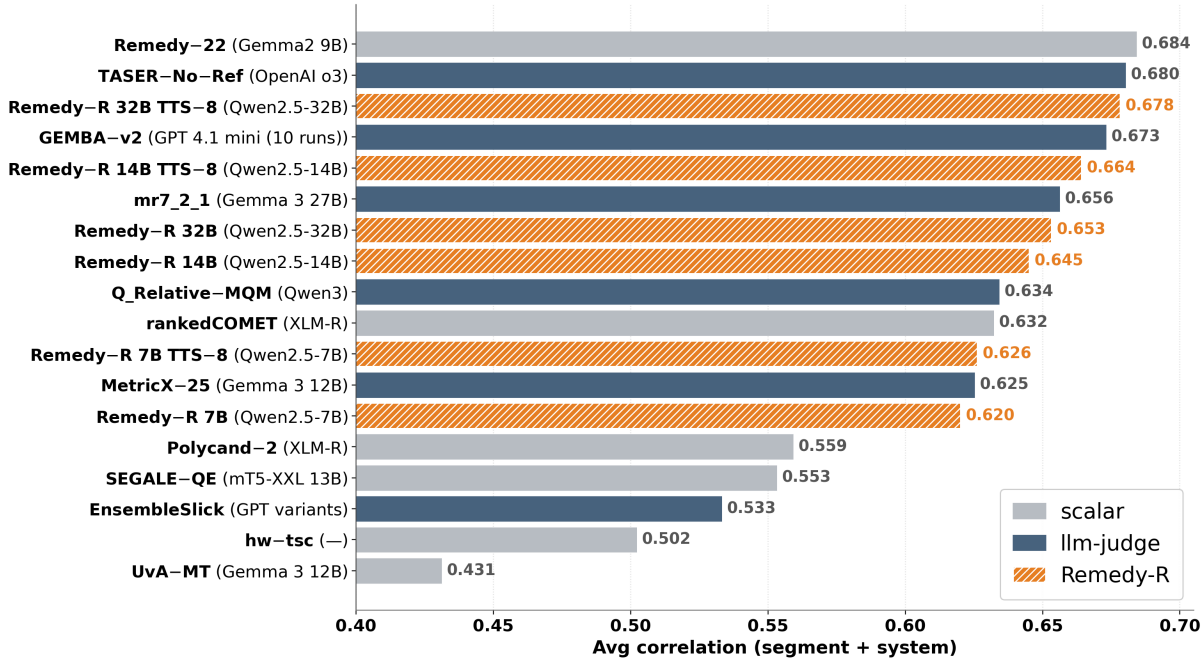

Figure 1: Average correlation performance on the WMT25 MQM. Following the official shared task protocol, we report the average correlation score combining system-level accuracy (**SPA**) and segment-level $\boldsymbol{acc}^*_{eq}$, using the *mt-metrics-eval* toolkit. We include WMT25 official submissions and their backbone models. Test-Time-Scaling (TTS) consistently improves correlation as the number of evaluation passes increases using Remedy-R.

ries at test time consistently enhances performance across all model sizes (see ablations on WMT24 in Figure 4). Notably, Remedy-R-14B reaches an average correlation of 74.9, matching the strongest GEMBA-MQM performance. The steady improvement from 7B to 32B suggests that iterative reasoning stabilizes evaluation outcomes and reduces stochastic variance, yielding more robust and reliable quality assessments.

Interestingly, we observe that TTS primarily improves *segment-level* acc^*_{eq} rather than system-level correlation (see Table 11). We hypothesize that this phenomenon is mostly due to the limitations of current meta-evaluation metrics. As noted by Perrella et al. (2024), tie-calibrated pairwise accuracy (acc^*_{eq}) tends to favor metrics that output continuous rather than discrete scores. Averaging multiple predictions effectively smooths discrete outputs into continuous scores, improving agreement with tie-calibrated accuracy, which favors metrics with finer score granularity.

Lastly, TTS remains effective on WMT25 (Figure 1), and we quantify the TTS variance analysis of the scores under fixed inputs in Table 6 and Appendix A.3 on WMT25, showing that repeated evaluation introduces substantially larger variability at the segment level than at the system level.

4.3 Analyses on Challenge sets

We evaluate Remedy-R under out-of-distribution (OOD) and adversarial conditions on MSLC24, and additionally construct an *unrelated translation* category by sampling 50 target-language sentences from Flores-200 (Costa-Jussà et al., 2022). For most perturbations (empty outputs, source copies, wrong-language, unrelated MT), a reliable metric should assign near-zero scores, while mixed-language outputs with correct meaning but code-switching should receive moderate scores.

Table 2 shows that scalar metrics (e.g., COMET-22, MetricX-24) can behave inconsistently under OOD inputs, sometimes assigning high scores to degenerate cases; for example, XCOMET scores 73.8%, 64.1%, and 82.0% on empty translation, empty source/reference, and source copy, respectively. In contrast, Remedy-R is robust and well-calibrated: it outputs near-zero scores for empty translations, sharply penalizes source-copy and wrong-language cases, and assigns moderate scores to mixed-language inputs.

Unlike MetricX-24 and XCOMET, which use synthetic augmentation that includes empty or hallucinated translations, Remedy-R is trained without such data, indicating stronger generalization beyond standard WMT conditions.

	ref?	empty mt	empty src+ref	src copy	wrong lang	mix lang	unrelated mt
COMET-22	✓	57.00%	58.81%	69.85%	67.84%	65.56%	45.23%
KIWI	✗	54.87%	67.72%	52.15%	82.64%	78.75%	41.95%
XCOMET	✓	73.79%	64.12%	82.04%	85.65%	71.77%	20.31%
MetricX-24-XXL	✓	-9.59	-5.85	-12.59	-3.06	-10.08	-24.15
MetricX-24-XXL	✗	-7.34	-5.85	-11.36	-2.51	-7.78	-24.25
GEMBA-ESA	✗	14.00%	13.5%	11.12%	14.32%	18.08%	1.27%
Remedy-R-7B	✓	1.00%	7.07%	76.92%	43.69%	60.6%	1.5%
Remedy-R-14B	✓	0.00%	0.00%	11.35%	14.6%	37.6%	0.6%
Remedy-R-32B	✓	0.00%	0.00%	2.76%	8.30%	46.0%	1.3%

Table 2: Averaged quality scores of different metric models on MSLC24 OOD set. For all classes except *mix-lang*, a robust metric should output low scores; for *mix-lang*, the translation preserves the source meaning but contains code-switching, so its quality scores should be moderately high rather than near zero. MetricX scores are ranged from -25 to 0 (higher is better). We provide additional reference-free Remedy-R results in Table 8.

5 Remedy-R Agent

So far, our analyses have focused on evaluation at the final score level, i.e., how well Remedy-R's predicted quality scores align with human ratings. However, such correlations do not necessarily imply that the generated analyses are faithful or practically useful.

We therefore evaluate Remedy-R's analyses in two complementary ways: (i) we prompt GPT-4o-mini to score explanation faithfulness given only the source and the translation (§5.1); and (ii) we reuse the analyses as feedback in a simple training-free evaluate–revise loop (*Remedy-R Agent*) and measure translation improvements (§5.2.1–§5.2.2).

5.1 How faithful are Remedy-R's evaluation explanations?

Remedy-R	en-de	en-ru	zh-en	Avg
7B	79.10	76.57	74.95	76.87
14B	80.02	76.75	78.05	78.27
32B	**81.18**	**79.05**	**78.38**	**79.54**

Table 3: GPT-4o-mini faithfulness scores for Remedy-R explanations on WMT22 MQM test samples (300 per language pair). GPT-4o-mini is given the source, translation, and explanation. Higher score means more faithful explanations.

We first assess explanation faithfulness by prompting GPT-4o-mini to score how well Remedy-R's reasoning is supported by the source sentence and the translation hypothesis only. We conduct this analysis on the WMT22 metric evaluation test set and sample 300 examples for each MQM language pair (en-de, en-ru, zh-en), resulting in 900 examples in total. Table 3 reports the average faithfulness scores. Overall, Remedy-R explanations receive consistently high faithfulness scores, and faithfulness improves with model scale.

5.2 Reusing Remedy-R Analyses for Training-Free Refinement

In the evaluate–revise setting, three models are involved:

- M_{base}: a translator that produces an initial translation for a source sentence;
- $M_{feedback}$: an evaluator that inspects the translation and generates an analysis as feedback;
- $M_{refinement}$: a refiner that revises the translation conditioned on (*src*, *mt*, *feedback*).

Instantiating these roles yields a training-free loop (*Remedy-R Agent*) that reuses Remedy-R's analyses for translation refinement. We evaluate the loop under controlled configurations and on strong open (Table 4 and 9) and closed (Figure 2) MT systems.

5.2.1 Controlled agent experiments with Qwen translators

We first study whether Remedy-R's analyses provide actionable feedback for translation refinement under a controlled setup. To isolate the effect of the feedback itself, we use Qwen2.5-Instruct models (7B, 14B, and 32B) as base translators and compare the following feedback–refinement configurations on WMT24, measured by XCOMET-XXL.

$M_{feedback}$	$M_{refinement}$	cs-uk	en-cs	en-de	en-es	en-hi	en-is	en-ja	en-ru	en-uk	en-zh	ja-zh	Avg
						M_{base} = Qwen2.5-it-7B | Remedy-R = 7B | x-Tower = 14B (w. XComet-XL = 3.5B)							
-	-	62.9	53.1	86.0	81.9	37.9	29.8	69.7	71.2	51.5	82.8	69.9	63.4
-	Base	64.3 ↑1.4	55.0 ↑1.9	86.4 ↑0.4	82.1 ↑0.2	39.1 ↑1.2	30.7 ↑0.9	69.6 ↓0.1	72.1 ↑0.9	53.8 ↑2.3	82.7 ↓0.1	70.4 ↑0.5	64.2 ↑0.8
x-Tower	Base	66.3 ↑3.4	55.8 ↑2.6	87.3 ↑1.3	83.0 ↑1.1	39.4 ↑1.5	30.3 ↑0.5	70.4 ↑0.7	74.2 ↑2.9	56.0 ↑4.4	83.0 ↑0.2	69.7 ↓0.2	65.0 ↑1.7
Remedy-R	Base	65.8 ↑2.9	55.9 ↑2.8	86.9 ↑0.9	82.9 ↑1.0	40.4 ↑2.5	30.5 ↑0.7	71.5 ↑1.8	73.3 ↑2.1	54.4 ↑2.9	83.0 ↑0.3	69.3 ↓0.6	64.9 ↑1.6
x-Tower	x-Tower	75.6 ↑12.6	57.0 ↑3.9	90.7 ↑4.7	85.8 ↑3.9	40.3 ↑2.4	33.6 ↑3.7	62.3 ↓7.4	79.3 ↑8.0	69.2 ↑17.7	82.2 ↓0.6	66.5 ↓3.4	67.5 ↑4.2
Remedy-R	Remedy-R	66.0 ↑3.1	56.0 ↑2.9	87.2 ↑1.1	83.2 ↑1.3	40.5 ↑2.6	30.9 ↑1.1	71.8 ↑2.1	73.4 ↑2.1	54.6 ↑3.1	83.2 ↑0.5	69.7 ↓0.2	65.1 ↑1.8
						M_{base} = Qwen2.5-it-14B | Remedy-R = 14B | x-Tower = 14B (w. XComet-XL = 3.5B)							
-	-	69.4	63.6	88.4	83.7	47.6	32.6	74.8	75.7	58.6	83.9	72.5	68.2
-	Base	71.0 ↑0.6	67.2 ↑3.6	89.4 ↑1.0	85.0 ↑1.3	51.5 ↑3.9	34.4 ↑1.9	77.9 ↑3.1	77.9 ↑2.2	63.2 ↑4.6	84.5 ↑0.6	72.7 ↑0.2	70.4 ↑2.2
x-Tower	Base	73.0 ↑3.6	68.5 ↑4.9	90.2 ↑1.8	85.3 ↑1.7	51.3 ↑3.7	34.6 ↑2.0	78.6 ↑3.8	78.5 ↑2.8	65.1 ↑6.5	84.3 ↑0.4	71.9 ↓0.6	71.0 ↑2.8
Remedy-R	Base	74.8 ↑5.4	68.2 ↑4.6	90.3 ↑1.9	85.6 ↑1.9	53.1 ↑5.5	36.3 ↑3.7	79.3 ↑4.5	79.2 ↑3.5	64.5 ↑5.9	84.1 ↑0.2	72.4 ↓0.1	71.6 ↑3.4
x-Tower	x-Tower	77.1 ↑7.7	62.5 ↓1.1	91.4 ↑3.0	86.5 ↑2.8	45.0 ↓2.6	34.1 ↑1.6	65.6 ↓9.2	80.2 ↑4.5	70.4 ↑11.8	82.8 ↓1.1	68.3 ↓4.2	69.4 ↑1.2
Remedy-R	Remedy-R	74.1 ↑4.7	68.3 ↑4.7	89.8 ↑1.4	84.9 ↑1.2	52.6 ↑5.0	35.8 ↑3.2	77.8 ↑2.9	77.4 ↑1.7	64.3 ↑5.7	83.8 ↓0.1	72.4 ↓0.1	71.0 ↑2.8
						M_{base} = Qwen2.5-it-32B | Remedy-R = 32B | x-Tower = 14B (w. XComet-XL = 3.5B)							
-	-	74.0	68.2	89.6	84.6	53.4	35.1	78.0	76.9	64.0	83.5	72.6	70.9
-	Base	75.0 ↑1.0	69.5 ↑1.3	90.0 ↑0.4	85.5 ↑0.9	55.4 ↑2.0	36.5 ↑1.4	78.7 ↑0.7	77.2 ↑0.3	66.7 ↑2.7	83.8 ↑0.3	73.6 ↑1.0	72.0 ↑1.1
x-Tower	Base	75.6 ↑1.6	70.4 ↑2.2	91.2 ↑1.6	86.1 ↑1.5	56.1 ↑2.7	36.5 ↑1.4	80.5 ↑2.6	80.3 ↑3.3	69.1 ↑5.1	84.8 ↑1.3	72.7 ↑0.1	73.0 ↑2.1
Remedy-R	Base	76.9 ↑2.9	71.5 ↑3.3	91.0 ↑1.4	85.9 ↑1.3	56.8 ↑3.4	37.2 ↑2.1	81.1 ↑3.2	79.4 ↑2.5	67.8 ↑3.8	83.9 ↑0.5	73.4 ↑0.8	73.2 ↑2.3
x-Tower	x-Tower	77.5 ↑3.5	63.5 ↓4.7	91.7 ↑2.1	86.1 ↑1.5	45.5 ↓7.9	34.2 ↓0.9	66.7 ↓11.2	80.2 ↑3.2	71.6 ↑7.6	82.5 ↓1.0	68.2 ↓4.5	69.8 ↓1.1
Remedy-R	Remedy-R	76.6 ↑2.6	71.5 ↑3.3	91.1 ↑1.5	85.6 ↑1.0	57.1 ↑3.7	37.4 ↑2.2	80.4 ↑2.5	78.5 ↑1.6	67.8 ↑3.8	84.0 ↑0.6	73.3 ↑0.7	73.0 ↑2.1

Table 4: Agent MT experiments on WMT24 using Qwen2.5 series models as the initial base translators (M_{base}, gray rows). We compare self-refinement without external feedback, feedback-guided refinement with xTower or Remedy-R, and unified settings where the same model is used for both feedback generation and refinement. We report translation quality with XCOMET-XXL. Note that the M_{base}=7B xTower→xTower result is not a size-matched comparison, as xTower is a 14B model.

We include x-Tower (Treviso et al., 2024), a GPT-4–distilled evaluator trained with span-level supervision from xCOMET-XL (Guerreiro et al., 2024), as the strongest explanation-based baseline.

- **Self-refinement** ($M_{feedback} = -$, $M_{refinement} = Base$): the Qwen translator revises its own output without external feedback.
- **xTower-Feedback** ($M_{feedback} = xTower$, $M_{refinement} = Base$): the Qwen translator revises its output using xTower's analysis as additional context.
- **Remedy-R-Feedback** ($M_{feedback} = Remedy\text{-}R$, $M_{refinement} = Base$): the Qwen translator revises its output using Remedy-R's analysis as additional context.

This controlled comparison allows us to isolate the value of the feedback while keeping the refiner fixed. Across all three model scales, adding external feedback improves over plain self-refinement, and Remedy-R feedback generally yields larger and more stable gains than xTower feedback. For example, under the 14B configuration, Remedy-R feedback improves the average XCOMET score by +3.4, compared with +2.8 for xTower feedback. These results suggest that Remedy-R's analyses are not merely descriptive, but provide actionable revision signals for translation improvement.

In addition, we also evaluate a unified configuration where Remedy-R is used both to generate feedback and to perform the refinement itself. This setting remains competitive with using Remedy-R only as the feedback model, indicating that the feedback can be directly reused for revision without any dedicated post-editing training.

Finally, we observe consistent improvements when applying Remedy-R Agent to ALMA-R (Xu et al., 2024a), further suggesting that the usefulness of Remedy-R's feedback is not limited to the Qwen family (Table 9 and Appendix A.7).

5.2.2 Applying Remedy-R Agent to strong closed LLMs

Finally, we test whether Remedy-R Agent remains effective when applied to strong closed LLM translators. Following prior work on multi-pass translation refinement (Briakou et al., 2024; Wu et al., 2025), we evaluate on the WMT24++ benchmark (Kocmi et al., 2024) and set M_{base} to GPT-4o-mini and Gemini-2.0-Flash. We use Remedy-R-32B as $M_{feedback}$ and apply iterative training-free evaluate–revise steps. For comparison, we include each model's internal self-refinement baselines, including step-by-step (Briakou et al., 2024) and translate-again (Wu et al., 2025).

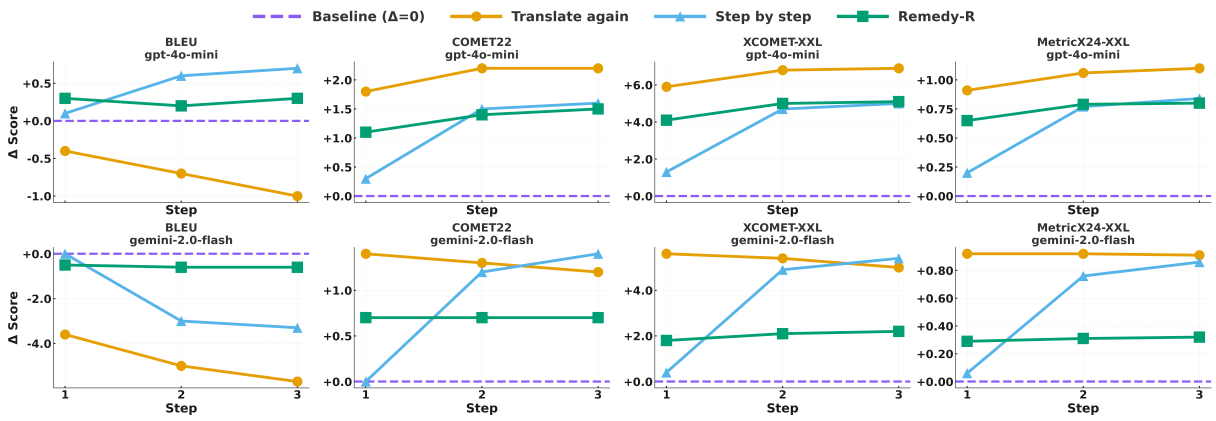

Figure 2: Refinement performance comparison on the initial translations from GPT-4o-mini and Gemini-2.0-Flash using paragraph-level WMT24++ benchmark. We use ref-based XCOMET-XXL to measure the translation quality. Here Translate-again and Step-by-Step utilize self-refinement (using GPT-4o-mini and gemini-2.0-flash themselves for refinement), while we use the 32B version for Remedy-R. More detailed results are in Appendix (Table 10).

As shown in Figure 2, Remedy-R Agent consistently improves translation quality on both closed LLMs. On GPT-4o-mini, Remedy-R Agent achieves gains comparable to step-by-step refinement, while on Gemini-2.0-Flash it attains a substantial fraction of the gains from self-refinement. Notably, Remedy-R Agent uses only a 32B open model for feedback, yet remains effective on translations produced by significantly larger closed LLMs, highlighting the generality of Remedy-R's feedback across model families.

6 Related Work

6.1 Automatic Translation Evaluation

Early MT metrics like BLEU (Papineni et al., 2002) and ChrF (Popović, 2015) rely on surface matching and correlate weakly with human judgments (Freitag et al., 2022). Black-box scalar metrics such as COMET (Rei et al., 2020) and Metric-X (Juraska et al., 2023) rely on regression over noisy human ratings, whereas Remedy (Tan and Monz, 2025) learns translation quality via reward modeling using pairwise preference data, achieving SOTA performance. For interpretability, xCOMET (Guerreiro et al., 2024) predicts MQM-based error spans, and xTower (Treviso et al., 2024) is trained to provide explanations by distilling GPT-4.

6.2 Reasoning for Machine Translation

Reasoning has been explored for both MT evaluation and generation. For evaluation, recent LLM-as-Judge systems use multi-stage or multi-agent designs to structure assessments (Feng et al., 2025b; Zhang et al., 2025). For generation, RL-based approaches optimize translation quality using learned rewards, e.g., MT-R1-Zero (Feng et al., 2025a) and Hunyuan-MT (Zheng et al., 2025).

6.3 Translation Agents

Agentic translation methods iteratively improve translations via self-refinement or decomposed sub-tasks, e.g., LLM-Refine (Xu et al., 2024c), translate-again (Wu et al., 2025) and step-by-step (Briakou et al., 2024). Closest to our setting, xTower (Treviso et al., 2024) can be used to support post-editing via span-based explanations. Our goal is not to propose a new translation pipeline; we use a simple evaluate–revise loop as a lightweight diagnostic to test whether Remedy-R's feedback can guide refinement.

7 Conclusions

A longstanding challenge in MT evaluation is limited explainability, which undermines metric reliability and robustness. We introduce Remedy-R, a generative reasoning-based MT metric trained with RLVR from pairwise human preferences using a verifiable reward, without error-span annotations. Remedy-R produces multi-dimensional analyses (accuracy, fluency, completeness) followed by a final score, achieves competitive performance on WMT22–25 benchmarks, and remains robust on OOD tests. Beyond score-level correlation, we validate the faithfulness and practical value of its quality analyses via Remedy-R Agent: a training-free evaluate–revise loop that consistently improves translations across models and language pairs, including strong open and commercial systems.

Limitations

In this paper, we do not conduct human evaluation to measure the faithfulness of Remedy-R's evaluation explanations due to limited budgets. However, we conduct these experiments with GPT-4o-mini with 900 random samples in the official WMT22 metric test set. As a generative evaluator, Remedy-R produces natural-language reasoning explanations that may not always perfectly reflect the underlying scoring process. We therefore include faithfulness checks and qualitative analyses; nonetheless, fully guaranteeing reasoning faithfulness remains an open challenge for all reasoning LLMs.

Broader Impact

We acknowledge there might be several ethical considerations in MT evaluation research such as gender bias. We prioritize high-quality open-source data and models in this research. We acknowledge that automatic quality estimation metrics can misguide Machine Translation in training and inference (Kocmi et al., 2025; Tan et al., 2025), which is the motivation we build Remedy-R toward more explainable, robust, and actionable Quality Estimation.

Acknowledgments

Christof Monz acknowledges funding by the Netherlands Organization for Scientific Research (NWO) under project number VI.C.192.080 and 2023.017. Shaomu Tan and Christof Monz thank SURF (www.surf.nl) for the support in using the Dutch National Supercomputer Snellius.

Shaomu Tan thanks Prof. Taro Watanabe and Prof. Hidetaka Kamigaito (Nara Institute of Science and Technology), Prof. Yoshimasa Tsuruoka and Zhongtao Miao (University of Tokyo), and Pinzhen Chen (University of Edinburgh) for their valuable feedback. Shaomu Tan also thanks Jiaqi Bao (Sony) for her moral support during his internship at Sony.

References

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

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 the Ninth Conference on Machine Translation*, pages 1301–1317.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

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

Daniel Deutsch, Rotem Dror, and Dan Roth. 2021. A statistical analysis of summarization evaluation metrics using resampling methods. *Transactions of the Association for Computational Linguistics*, 9:1132–1146.

Daniel Deutsch, George Foster, and Markus Freitag. 2023. Ties matter: Meta-evaluating modern metrics with pairwise accuracy and tie calibration. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12914–12929.

Zhaopeng Feng, Shaosheng Cao, Jiahan Ren, Jiayuan Su, Ruizhe Chen, Yan Zhang, Zhe Xu, Yao Hu, Jian Wu, and Zuozhu Liu. 2025a. Mt-r1-zero: Advancing llm-based machine translation via r1-zero-like reinforcement learning. *arXiv preprint arXiv:2504.10160*.

Zhaopeng Feng, Jiayuan Su, Jiamei Zheng, Jiahan Ren, Yan Zhang, Jian Wu, Hongwei Wang, and Zuozhu Liu. 2025b. M-MAD: Multidimensional multi-agent debate for advanced machine translation evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7084–7107.

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

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Nuno M Guerreiro, Ricardo Rei, Daan Van Stigt, Luísa Coheur, Pierre Colombo, and André FT Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. 2025. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. *arXiv preprint arXiv:2503.24290*.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. Metricx-24: The google submission to the wmt 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. Metricx-23: The google submission to the wmt 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767.

Rebecca Knowles, Samuel Larkin, and Chi-Kiu Lo. 2024. Mslc24: Further challenges for metrics on a wide landscape of translation quality. In *Proceedings of the Ninth Conference on Machine Translation*, pages 475–491.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, et al. 2025. Findings of the wmt25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, et al. 2024. Findings of the wmt24 general machine translation shared task: The llm era is here but mt is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46.

Tom Kocmi and Christian Federmann. 2023a. Gemba-mqm: Detecting translation quality error spans with gpt-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775.

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

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494.

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 *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

Alon Lavie, Greg Hanneman, Sweta Agrawal, Diptesh Kanojia, Chi-Kiu Lo, Vilém Zouhar, Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, et al. 2025. Findings of the wmt25 shared task on automated translation evaluation systems: Linguistic diversity is challenging and references still help. In *Proceedings of the Tenth Conference on Machine Translation*, pages 436–483.

Chi-kiu Lo, Samuel Larkin, and Rebecca Knowles. 2023. Metric score landscape challenge (mslc23): Understanding metrics' performance on a wider landscape of translation quality. In *Proceedings of the Eighth Conference on Machine Translation*, pages 776–799.

Qingyu Lu, Liang Ding, Kanjian Zhang, Jinxia Zhang, and Dacheng Tao. 2025. Mqm-ape: Toward high-quality error annotation predictors with automatic post-editing in llm translation evaluators. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5570–5587.

Qingyu Lu, Baopu Qiu, Liang Ding, Kanjian Zhang, Tom Kocmi, and Dacheng Tao. 2024. Error analysis prompting enables human-like translation evaluation in large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 8801–8816.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020. Results of the wmt20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725.

Nikita Moghe, Arnisa Fazla, Chantal Amrhein, Tom Kocmi, Mark Steedman, Alexandra Birch, Rico Sennrich, and Liane Guillou. 2025. Machine translation meta evaluation through translation accuracy challenge sets. *Computational Linguistics*, 51(1):73–137.

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

Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Edoardo Barba, and Roberto Navigli. 2024. Guardians of the machine translation meta-evaluation: Sentinel metrics fall in! In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16216–16244.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Lorenzo Proietti, Stefano Perrella, and Roberto Navigli. 2025. Has machine translation evaluation achieved human parity? the human reference and the limits of progress. *arXiv preprint arXiv:2506.19571*.

Shenbin Qian, Archchana Sindhujan, Minnie Kabra, Diptesh Kanojia, Constantin Orašan, Tharindu Ranasinghe, and Fred Blain. 2024. What do large language models need for machine translation evaluation? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3660–3674.

Ricardo Rei, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT Martins. 2022. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585.

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. *arXiv preprint arXiv:1707.06347*.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2024. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*.

Richard Sproat, Tianyu Zhao, and Llion Jones. 2025. Transevalnia: Reasoning-based evaluation and ranking of translations. *arXiv preprint arXiv:2507.12724*.

Shaomu Tan, Ryosuke Mitani, Ritvik Choudhary, and Toshiyuki Sekiya. 2025. Investigating test-time scaling with reranking for machine translation. *arXiv preprint arXiv:2509.19020*.

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*, pages 4370–4387, Suzhou, China. Association for Computational Linguistics.

Brian Thompson, Nitika Mathur, Daniel Deutsch, and Huda Khayrallah. 2024. Improving statistical significance in human evaluation of automatic metrics via soft pairwise accuracy. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1222–1234.

Marcos Treviso, Nuno Guerreiro, Sweta Agrawal, Ricardo Rei, José Pombal, Tânia Vaz, Helena Wu, Beatriz Silva, Daan Stigt, and André FT Martins. 2024. xtower: A multilingual llm for explaining and correcting translation errors. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15222–15239.

Di Wu, Seth Aycock, and Christof Monz. 2025. Please translate again: Two simple experiments on whether human-like reasoning helps translation. *arXiv preprint arXiv:2506.04521*.

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 *The Twelfth International Conference on Learning Representations*.

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 *International Conference on Machine Learning*, pages 55204–55224. PMLR.

Wenda Xu, Daniel Deutsch, Mara Finkelstein, Juraj Juraska, Biao Zhang, Zhongtao Liu, William Yang Wang, Lei Li, and Markus Freitag. 2024c. Llmrefine: Pinpointing and refining large language models via fine-grained actionable feedback. In *Findings of the*

Association for Computational Linguistics: NAACL 2024, pages 1429–1445.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yi-Chao Zhang, Yunyang Wan, Yuqi Liu, Zeyu Cui, Zhenru Zhang, Zihan Qiu, Shanghaoran Quan, and Zekun Wang. 2024. Qwen2.5 technical report. *ArXiv*, abs/2412.15115.

Shijie Zhang, Renhao Li, Songsheng Wang, Philipp Koehn, Min Yang, and Derek F Wong. 2025. Himate: A hierarchical multi-agent framework for machine translation evaluation. *arXiv preprint arXiv:2505.16281*.

Mao Zheng, Zheng Li, Bingxin Qu, Mingyang Song, Yang Du, Mingrui Sun, and Di Wang. 2025. Hunyuan-mt technical report. *arXiv preprint arXiv:2509.05209*.

A Appendix

A.1 RLVR Training Details

We train Remedy-R using reinforcement learning with the verifiable reward in §2.2. The model acts as a policy that, given an input, generates reasoning steps and final scores, and then receives a scalar reward that reflects alignment with human judgments. This objective is to update the model so that high-reward behaviors become more likely over time. Formally, the model defines a conditional policy $\pi_\theta(\mathbf{y} \mid \mathbf{x})$ over output sequences. After generating a response, we compute a scalar reward and maximize the expected return with gradients estimated via the policy gradient theorem:

$$\begin{aligned} \mathcal{J}(\theta) &= \mathbb{E}_{\mathbf{x},\mathbf{y}\sim\pi_\theta}\big[r(\mathbf{y},\mathbf{x})\big], \\ \nabla_\theta \mathcal{J}(\theta) &= \mathbb{E}\Big[\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta\big(y_t \mid y_{<t},\mathbf{x}\big)\, A_t\Big]. \end{aligned} \tag{6}$$

where A_t denotes the token-level advantage. This estimator increases the likelihood of high-reward generations and decreases that of low-reward ones, aligning the model's behavior with the reward signal. We optimize this objective using Proximal Policy Optimization (PPO) (Schulman et al., 2017), which stabilizes updates by clipping the ratio between the new and the old policy, while regularizing against a frozen reference model $\pi_{\theta_{\text{ref}}}$ to prevent reward over-optimization. During training, each rollout prompts the model with a translation pair, generates an evaluation output, parses the scores, computes the reward, estimates token level advantages, and finally applies PPO updates. We then optimize the following PPO objective:

$$\begin{aligned} \mathcal{L}_{\text{PPO}}(\theta) &= \mathbb{E}_t[\ell_t(\theta)] - \beta_{\text{KL}}\,\text{KL}(\pi_\theta \,\|\, \pi_{\text{ref}})\,, \\ \ell_t(\theta) &= \min\Big(r_t(\theta)A_t,\ \text{clip}_\epsilon\big(r_t(\theta)\big)A_t\Big). \end{aligned} \tag{7}$$

Here, $r_t(\theta) = \frac{\pi_\theta(y_t|y_{<t},\mathbf{x})}{\pi_{\theta_{\text{old}}}(y_t|y_{<t},\mathbf{x})}$ is the likelihood ratio, ϵ is the clipping threshold, and β_{KL} controls the strength of the KL penalty that regularizes the updated policy toward a frozen reference policy $\pi_{\theta_{\text{ref}}}$ (the corresponding pretrained base model of the same size). Rewards are terminal and defined on the final parsed score. For advantage estimation, we use Generalized Advantage Estimation (GAE) with $\lambda = 1$, which trades off bias and variance by exponentially weighting multi step returns:

$$\begin{aligned} A_t &= \sum_{l=0}^{\infty} (\gamma\lambda)^l \delta_{t+l}, \\ \delta_t &= r_t + \gamma V_\phi(s_{t+1}) - V_\phi(s_t). \end{aligned} \tag{8}$$

where V_ϕ is the value function, and $\gamma \in (0, 1]$ and $\lambda \in [0, 1]$ are standard discounting hyperparameters. Following recent work (Hu et al., 2025), we set $\lambda = 1$, which simplifies the estimator to a discounted Monte Carlo return and improves stability in practice.

A.2 Ablation study on WMT22 Metric Benchmark

We conduct an ablation study on the WMT22 metric benchmark to isolate the effect of our reward design during RLVR training. Unless otherwise specified, all ablations in this subsection are performed with **Remedy-R 7B**. We compare two reward settings: (1) a pure pairwise ranking reward that only verifies whether the model-implied preference matches the human preference label; and (2) adding Huber reward shaping to encourage better calibration and reduce overly discrete score behaviors.

As shown in Table 5, adding Huber reward shaping consistently improves accuracy at both the system and segment levels. In particular, Huber shaping improves system-level accuracy from 87.6% to 89.1% and segment-level accuracy from 52.7% to 54.8%, yielding a +1.7% absolute gain in average accuracy. These results suggest that Huber shaping provides a more informative learning signal beyond binary preference verification and leads to better-aligned quality judgments.

Model	Reward Setting	Sys	Seg	Avg
7B	Pairwise Ranking Reward	87.6%	52.7%	70.2%
	+ Huber Reward Shaping	89.1%	54.8%	71.9%
14B	Pairwise + Huber	88.7%	56.0%	72.4%
	Pairwise + Huber + genRM penalty	89.8%	56.3%	73.1%

Table 5: Ablation study on WMT22 comparing reward designs. We report accuracy at the system and segment levels, and their average.

We further explore incorporating an explanation quality penalty into the reward. Instead of training a separate reward model, we reuse the same base instruct model (the RL initialization) as a generative rationale judge (genRM). Given the source, translation, and the model-generated explanation, genRM produces a scalar score intended to reflect

Remedy-R Training Prompt Template

You are an expert machine translation evaluator. You need to assess the quality of two translations of the same source text. Your task is to evaluate the translation quality and provide scores from 0–100, where higher scores indicate better quality. You are also given a reference (not always perfect) to help you assess the quality.

Evaluation Criteria:

- Accuracy: Whether the meaning expressed in the translation is correct and faithful to the source. Penalize mistranslation, unsupported additions/hallucinations, terminology errors, and untranslated text.
- Fluency: How natural and grammatically correct the translation reads in the target language. Consider grammar, agreement, word order, punctuation, spelling, register.
- Completeness: Is all information from the source conveyed without omissions?

Instructions: Think step by step about the quality of each translation and write your analysis first, then provide your final scores. Evaluate each translation independently rather than by comparison.

Output Format: Thinking through your evaluation first, then output the scores in exactly this format (do not give scores first):

A: [score] B: [score]

Now evaluate this $SRC_LANG–$TGT_LANG translation:

—

Source: $SOURCE
Reference: $REFERENCE
Translation A: $TRANSLATION_A
Translation B: $TRANSLATION_B

the rationale's faithfulness and relevance. We subtract this score (scaled by a fixed coefficient) from the RL reward, penalizing low-quality explanations while keeping the pairwise preference verification term unchanged.

We provide the reward dynamics during training in Figure 3. As shown in Table 5, adding this penalty results in only a marginal change in accuracy in our Remedy-R 14B setting. Overall, this suggests that the main gains come from the pairwise reward with Huber shaping, while explanation-based regularization provides at most a small additional benefit under our current design.

A.3 Test Time Scaling with multiple Evaluation Passes

Remedy-R's generative reasoning nature enables the application of **Test-Time Scaling (TTS)**, where multiple evaluation passes are performed with different reasoning trajectories and their quality scores are aggregated. In this setting, we adopt a simple implementation that averages the quality scores from multiple independent evaluations.

As shown in Figure 4, performing more evaluation trajectories at test time consistently enhances performance across all model sizes. Notably, Remedy-R-14B reaches an average correlation of 74.9, matching the strongest GEMBA-MQM performance. The steady improvement from 7B to 32B suggests that iterative reasoning stabilizes evaluation outcomes and reduces stochastic variance, yielding more robust and reliable quality assessments.

Interestingly, we observe that TTS primarily improves *segment-level* acc^*_{eq} rather than system-level correlation (see Table 11). We hypothesize that this phenomenon is mostly due to the limitations of current meta-evaluation metrics. As noted by Perrella et al. (2024), tie-calibrated pairwise accuracy (acc^*_{eq}) tends to favor metrics that output continuous rather than discrete scores. Averaging multiple predictions effectively smooths discrete outputs into continuous scores, improving agreement with tie-calibrated accuracy, which favors metrics with finer score granularity.

Temp	Level	std_{mean}	std_{p90}	$\text{range}_{\text{mean}}$	$\text{range}_{\text{p90}}$
0.2	Segment	0.91	2.42	2.24	5.00
0.2	System	0.121	0.167	0.382	0.546
1.0	Segment	1.48	2.50	3.75	5.00
1.0	System	0.202	0.272	0.630	0.886

Table 6: Fixed-input variance analysis of Remedy-R under repeated evaluation (TTS-8) on the WMT25 MQM subset. We report the mean and 90th percentile of the standard deviation and min–max range across 8 repeated evaluations. Variability is substantially lower at the system level than at the segment level, showing that aggregation improves reproducibility of final system-level scores.

Variance analysis for Test-Time Scaling. To further examine the stochasticity of Remedy-R under Test-Time Scaling (TTS), we conduct a fixed-input variance analysis on the WMT25 MQM subset

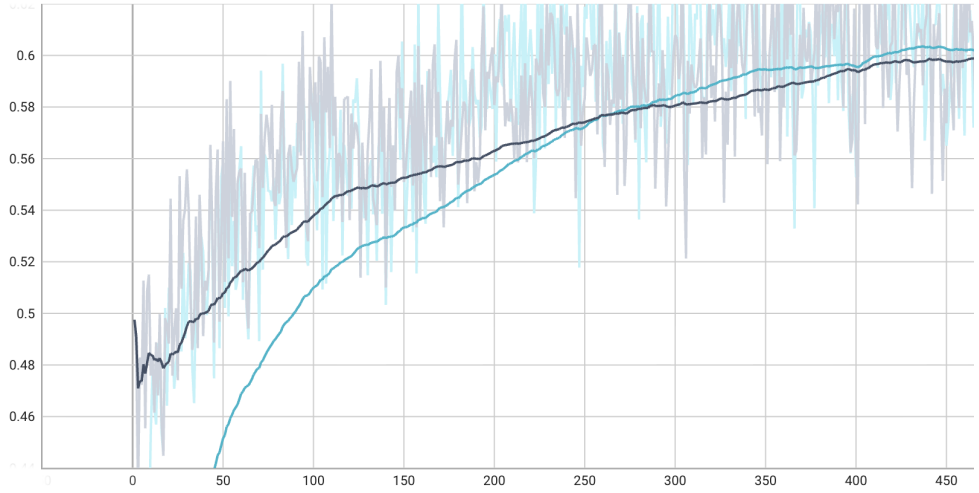

Figure 3: Reward curves during training for Pairwise + Huber (dark-blue) and Pairwise + Huber + genRM penalty (light-blue) settings.

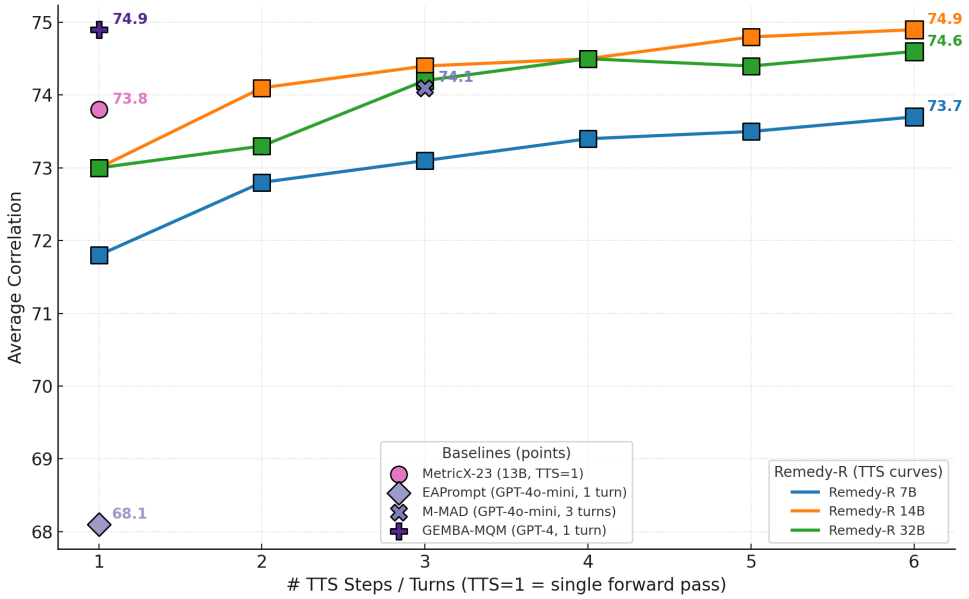

Figure 4: Average correlation across WMT23 MQM benchmarks under different numbers of Test-Time Scaling (TTS) evaluation passes. Each configuration aggregates multiple independent evaluations by averaging their final quality scores. TTS consistently improves correlation as the number of evaluation passes increases. Full results are shown in Table 11 in Appendix.

(5,242 segments, 39 systems) using the 32B model with 8 repeated evaluations per input (TTS-8).

For each segment, we compute the standard deviation and min–max range across 8 repeated evaluations. For each system, we first aggregate segment scores within each run to obtain one system score per run, and then compute variability across runs. Table 6 reports the mean and 90th percentile of these statistics at both the segment and system levels.

The results show two clear trends. First, stochasticity increases with temperature, as expected. Second, variability is substantially larger at the segment level than at the system level, indicating that aggregation across segments markedly improves the reproducibility of final system-level evaluation results. This supports our claim that repeated evaluation and score aggregation help stabilize Remedy-R's predictions under TTS.

A.4 WMT23 Metric Benchmark

We report results on the WMT23 MQM metric shared task in Table 11. We compare Remedy-R against established reference-based metrics (e.g.,

XCOMET-XXL, MetricX-23), QE variants, and recent LLM-based judges. For Remedy-R, we additionally evaluate test-time scaling (TTS) by sampling multiple reasoning outputs and aggregating their predicted scores (TTS=k). We report both system-level accuracy (**Acc**) and segment-level $\boldsymbol{acc}^*_{\boldsymbol{eq}}$, together with the average of the two.

As shown in Table 11, increasing TTS generally improves $\boldsymbol{acc}^*_{\boldsymbol{eq}}$ more consistently than **Acc**, leading to steady gains in the overall average. This trend is most pronounced for smaller models (7B), where TTS closes a substantial portion of the gap to stronger baselines, and remains beneficial for 14B and 32B as well. These results support our claim that sampling-based aggregation helps mitigate discrete score behaviors and yields more stable segment-level judgments.

A.5 WMT24 Metric Benchmark

Methods	Rank	Avg corr	Sys SPA	Seg $\boldsymbol{acc}^*_{\boldsymbol{eq}}$
Remedy$_{\text{9B-24}}$	1	**72.9**	85.9	**60.0**
Remedy-R-14B	2	72.6	**87.9**	57.2
MetricX-24-Hybrid	3	72.1	85.6	58.5
XCOMET-XXL	4	71.9	86.1	57.6
MetricX-24-Hybrid-QE	5	71.4	84.9	58.0
GEMBA-ESA (GPT4)	6	71.1	84.6	57.6
Remedy-R-32B	7	70.6	85.3	56.2
Remedy-R-7B	8	70.6	85.8	55.4
XCOMET-XXL-QE	9	69.5	83.3	55.7
Skywork-RM	10	69.0	83.2	54.7

Table 7: Evaluation on WMT24 MQM set. We report the official accuracy percentage (SPA and acc^*_{eq}).

Table 7 reports results on the WMT24 MQM metric benchmark. Following the official shared task protocol, we report system-level accuracy (**SPA**) and segment-level $\boldsymbol{acc}^*_{\boldsymbol{eq}}$, as well as their average correlation score (**Avg corr**) and the corresponding rank.

Remedy-R achieves competitive performance on WMT24 across model scales. In particular, Remedy-R-14B ranks second overall and attains the best system-level accuracy (**SPA**). While Remedy-R-32B and Remedy-R-7B show slightly lower segment-level $\boldsymbol{acc}^*_{\boldsymbol{eq}}$ than top baselines, their overall average remains comparable to strong reference-based metrics such as MetricX-24-Hybrid and XCOMET-XXL.

A.6 GPT-4o-mini Faithfulness Prompt

We use the following prompt to assess the faithfulness of Remedy-R's evaluation explanations. GPT-4 is given the source sentence, the translation hypothesis, and the explanation only (no reference translation), and returns a JSON object containing a `faithfulness_score` and a brief reason.

```
System:
You are a strict verifier. Your job is to score
the FAITHFULNESS of an evaluation explanation.
You must return ONLY a single valid JSON object
and nothing else.

User:
You are given: 1) src_sent: the source sentence;
2) target_sent: the translation hypothesis; 3)
explanation: an evaluation text that comments on
the translation quality

Task: Provide a score (0-100) indicating how
FAITHFUL the explanation is to src_sent and
target_sent.

Definition of faithfulness:
- Every key claim in the explanation must be
supported by what is actually present in src_sent
and/or target_sent.
- If the explanation invents content, mentions
errors that are not evidenced, misquotes words,
or contradicts src/target, score lower.

CRITICAL:
- You are NOT evaluating translation quality.
- A translation can be very bad, but an
explanation can still be highly faithful if it
correctly describes that badness.

Return ONLY JSON with:
{"faithfulness_score": int,   // 0-100
  "brief_reason": string       // <= 40 words,
  cite the biggest supported or unsupported
  claim}

Input:
src_lang: <src_lang>; tgt_lang: <tgt_lang>;
src_sent: <src>; target_sent: <tgt>; explanation:
<explanation>
```

A.7 Remedy-R Agent experiments

Table 9 reports additional Remedy-R Agent results on WMT24 using strong open-source ALMA-R (Xu et al., 2024b) translators as M_{base}. We evaluate two base settings: ALMA-R-7B with Remedy-R-7B, and ALMA-R-13B with Remedy-R-14B. For each language direction, we report multiple automatic metrics, including SacreBLEU, XCOMET-XXL, MetricX-24-XXL, and Remedy-R's own scores. We include both the initial translations (row "- / -") and the refined outputs produced by the Remedy-R Agent (row "Remedy-R

	ref?	empty mt	empty src+ref	src copy	wrong lang	mix lang	unrelated mt
COMET-22	✓	57.00%	58.81%	69.85%	67.84%	65.56%	45.23%
KIWI	✗	54.87%	67.72%	52.15%	82.64%	78.75%	41.95%
XCOMET	✓	73.79%	64.12%	82.04%	85.65%	71.77%	20.31%
MetricX-24-XXL	✓	-9.59	-5.85	-12.59	-3.06	-10.08	-24.15
MetricX-24-XXL	✗	-7.34	-5.85	-11.36	-2.51	-7.78	-24.25
GEMBA-ESA	✗	14.00%	13.5%	11.12%	14.32%	18.08%	1.27%
Remedy-R-7B	✓	1.00%	7.07%	76.92%	43.69%	60.6%	1.5%
Remedy-R-7B	✗	0.83%	5.40%	90.38%	33.15%	65.4%	2.0%
Remedy-R-14B	✓	0.00%	0.00%	11.35%	14.6%	37.6%	0.6%
Remedy-R-14B	✗	0.00%	0.00%	28.07%	12.1%	35.5%	1.0%
Remedy-R-32B	✓	0.00%	0.00%	2.76%	8.30%	46.0%	1.3%
Remedy-R-32B	✗	0.00%	0.00%	1.30%	8.0%	43.8%	3.5%

Table 8: Averaged quality scores of different metric models on MSLC24 OOD set. For all classes except *mix-lang*, a robust metric should output low scores; for *mix-lang*, the translation preserves the source meaning but contains code-switching, so its quality scores should be moderately high rather than near zero. MetricX scores are ranged from -25 to 0.

/ Remedy-R"), where Remedy-R generates the explanation and performs the refinement based on its own feedback.

Overall, Remedy-R Agent consistently improves translation quality across language pairs, with particularly large gains on en-zh. These results complement the main agent experiments and further demonstrate that Remedy-R's reasoning can be reused to drive refinement even when the initial translations come from a strong external MT system.

A.8 Remedy-R examples

We further provide several examples of the Remedy-R-14B model, including Chinese-English, English-Japanese, English-German, English-Czech, English-Spanish.

$M_{feedback}$	$M_{refinement}$	en-cs	en-de	en-ru	en-zh	Avg	$M_{feedback}$	$M_{refinement}$	en-cs	en-de	en-ru	en-zh	Avg
Base = ALMA-R-7B | Remedy-R = 7B													
SacreBLEU							**XCOMET-XXL**						
-	-	16.6	22.2	14.1	23.9	19.2	-	-	71.9	89.5	77.3	75.7	78.6
Remedy-R	Remedy-R	17.5 ↑0.9	22.2 ↑0	15.8 ↑1.7	27.4 ↑3.5	20.7 ↑1.5	Remedy-R	Remedy-R	71.4 ↓0.5	90.4 ↑0.9	78.5 ↑1.3	79.9 ↑4.2	80.0 ↑1.4
MetricX-24-XXL							**Remedy-R**						
-	-	-6.2	-2.5	-4.5	-3.5	-4.2	-	-	92.3	93.4	92.5	86.5	91.2
Remedy-R	Remedy-R	-6.3 ↑0.1	-2.4 ↑0.2	-4.2 ↑0.3	-3.0 ↑0.5	-4.0 ↑0.2	Remedy-R	Remedy-R	94.2 ↑1.9	95.4 ↑2.0	94.6 ↑2.1	92.0 ↑5.5	94.1 ↑2.9
Base = ALMA-R-13B | Remedy-R = 14B													
SacreBLEU							**XCOMET-XXL**						
-	-	18.2	22.8	15.9	26.2	20.8	-	-	76.5	91.0	80.4	79.6	81.9
Remedy-R	Remedy-R	20.1 ↑1.9	23.2 ↑0.4	17.3 ↑1.4	34.8 ↑8.6	23.8 ↑3.1	Remedy-R	Remedy-R	77.3 ↑0.8	91.4 ↑0.3	80.7 ↑0.2	84.0 ↑4.5	83.3 ↑1.4
MetricX-24-XXL							**Remedy-R**						
-	-	-5.2	-2.1	-3.8	-3.1	-3.6	-	-	90.5	91.4	90.5	88.7	90.3
Remedy-R	Remedy-R	-5.3 ↓0.1	-2.0 ↑0.1	-3.8 ↑0.1	-2.5 ↑0.5	-3.4 ↑0.2	Remedy-R	Remedy-R	92.1 ↑1.6	93.1 ↑1.7	91.9 ↑1.4	93.2 ↑4.5	92.6 ↑2.3

Table 9: Agent MT experiments on WMT24 using ALMA-R models as M_{base}. We report initial translations ("- / -") and Remedy-R Agent refinements ("Remedy-R / Remedy-R") under multiple metrics. Left: SacreBLEU. Right: XCOMET-XXL. We additionally report MetricX-24-XXL and Remedy-R scores in the lower blocks.

Method	Step	en-cs	en-de	en-fr	en-ja	en-ru	en-zh	Avg
gpt-4o-mini as base translator model								
gpt-4o-mini	base	54.5	78.6	61.7	58.7	58.0	57.3	61.5
translate_again	1	61.4 ↑6.9	81.9 ↑3.4	68.4 ↑6.6	64.3 ↑5.6	64.6 ↑6.6	63.5 ↑6.3	67.4 ↑5.9
translate_again	2	62.5 ↑7.9	82.0 ↑3.4	69.8 ↑8.1	65.6 ↑7.0	65.1 ↑7.1	64.6 ↑7.3	68.3 ↑6.8
translate_again	3	62.9 ↑8.4	82.5 ↑3.9	69.7 ↑7.9	65.5 ↑6.8	65.6 ↑7.6	64.5 ↑7.2	68.4 ↑7.0
step_by_step	1	56.6 ↑2.0	78.7 ↑0.1	64.2 ↑2.4	59.1 ↑0.5	59.8 ↑1.8	58.5 ↑1.2	62.8 ↑1.3
step_by_step	2	60.7 ↑6.2	80.8 ↑2.3	67.9 ↑6.2	62.5 ↑3.8	63.0 ↑5.0	62.5 ↑5.2	66.2 ↑4.8
step_by_step	3	61.2 ↑6.7	81.2 ↑2.6	68.2 ↑6.5	62.5 ↑3.8	63.6 ↑5.6	62.5 ↑5.2	66.5 ↑5.1
Remedy-R	1	57.6 ↑3.1	80.3 ↑1.7	66.3 ↑4.6	63.0 ↑4.4	63.2 ↑5.2	63.4 ↑6.2	65.6 ↑4.2
Remedy-R	2	58.0 ↑3.5	81.3 ↑2.7	67.2 ↑5.5	64.5 ↑5.8	64.1 ↑6.1	64.0 ↑6.7	66.5 ↑5.1
Remedy-R	3	58.4 ↑3.8	81.3 ↑2.7	67.3 ↑5.6	64.2 ↑5.5	64.3 ↑6.3	64.2 ↑7.0	66.6 ↑5.2
gemini-2.0-flash as base translator model								
gemini-2.0-flash	base	63.0	80.3	64.5	67.6	65.0	62.3	67.1
translate_again	1	68.7 ↑5.6	84.4 ↑4.1	70.8 ↑6.4	71.2 ↑3.6	71.1 ↑6.1	70.0 ↑7.7	72.7 ↑5.6
translate_again	2	68.5 ↑5.5	84.1 ↑3.8	69.6 ↑5.2	70.9 ↑3.3	71.2 ↑6.2	70.8 ↑8.4	72.5 ↑5.4
translate_again	3	68.7 ↑5.7	84.1 ↑3.8	68.6 ↑4.1	70.2 ↑2.6	70.7 ↑5.7	70.3 ↑7.9	72.1 ↑5.0
step_by_step	1	64.5 ↑1.4	81.0 ↑0.7	65.6 ↑1.2	66.3 ↓1.3	65.6 ↑0.6	62.0 ↓0.3	67.5 ↑0.4
step_by_step	2	68.9 ↑5.8	84.2 ↑3.8	69.7 ↑5.3	70.0 ↑2.4	70.8 ↑5.8	68.4 ↑6.1	72.0 ↑4.9
step_by_step	3	69.3 ↑6.2	84.5 ↑4.1	70.3 ↑5.8	70.4 ↑2.8	71.0 ↑6.0	69.5 ↑7.2	72.5 ↑5.4
Remedy-R	1	62.5 ↓0.5	82.0 ↑1.6	67.8 ↑3.3	69.6 ↑2.0	66.5 ↑1.5	65.1 ↑2.7	68.9 ↑1.8
Remedy-R	2	62.8 ↓0.2	82.3 ↑2.0	67.6 ↑3.1	69.9 ↑2.3	66.6 ↑1.6	65.9 ↑3.5	69.2 ↑2.0
Remedy-R	3	62.8 ↓0.3	81.9 ↑1.5	67.9 ↑3.4	69.9 ↑2.3	66.7 ↑1.6	65.8 ↑3.5	69.2 ↑2.0

Table 10: Refinement performance comparison on the initial translations from GPT-4o-mini and Gemini-2.0-Flash using paragraph-level WMT24++ benchmark. We report ref-based XCOMET-XXL to measure the translation quality. Step here means refinement steps or iterations. For translate_again and step_by_step, they adopt the self-refinement (using gpt-4o-mini and gemini-2.0-flash refinement), while Remedy-R utilizes 32B model Remedy-R model for refinement.

Method	θ	TTS/Turns	Acc	acc^*_{eq}	Avg
KIWI-XXL	ensemble	1	91.1	54.6	72.9
MetricX-23	13B	1	90.7	56.9	73.8
MetricX-23-QE	13B	1	89.0	56.1	72.6
XCOMET-XXL	ensemble	1	92.8	57.7	75.3
XCOMET-XXL-QE	ensemble	1	91.6	55.8	73.7
Remedy$_{\text{9B-23}}$	9B	1	94.1	58.2	76.2
EAPrompt (GPT-4o-mini)	-	1 turn	90.3	45.9	68.1
M-MAD (GPT-4o-mini)	-	3 turns	94.5	53.7	74.1
GEMBA-MQM (GPT4)	-	1 turn	94.5	55.2	74.9
Remedy-R	7B	1 turn	93.7	49.9	71.8
Remedy-R	7B	TTS=2	94.1	51.5	72.8
Remedy-R	7B	TTS=3	93.7	52.6	73.1
Remedy-R	7B	TTS=4	93.7	53.1	73.4
Remedy-R	7B	TTS=5	93.7	53.4	73.5
Remedy-R	7B	TTS=6	93.7	53.6	73.7
Remedy-R	14B	1 turn	92.4	53.6	73.0
Remedy-R	14B	TTS=2	93.3	55.0	74.1
Remedy-R	14B	TTS=3	93.7	55.1	74.4
Remedy-R	14B	TTS=4	93.7	55.4	74.5
Remedy-R	14B	TTS=5	94.1	55.4	74.8
Remedy-R	14B	TTS=6	94.1	55.7	74.9
Remedy-R	32B	1 turn	94.5	51.5	73.0
Remedy-R	32B	TTS=2	94.1	52.6	73.3
Remedy-R	32B	TTS=3	94.5	53.9	74.2
Remedy-R	32B	TTS=4	94.9	54.1	74.5
Remedy-R	32B	TTS=5	94.5	54.3	74.4
Remedy-R	32B	TTS=5	95.0	54.3	74.6

Table 11: Evaluation on WMT MQM23 Metric Shared task. Both KIWI-XXL and XCOMET-XXL are identical ensembles of 2 $\times$ 10.7B and 1 $\times$ 3.5B models.

Remedy-R (14B) ref-free Evaluation Example (Chinese-English)

Source: 麦乐鸡，没有麦乐鸡酱

Translation: Wheat chicken, no maillot chicken sauce

Remedy-R's Evaluation:
Let's evaluate the translation step by step based on the given criteria:

- **Accuracy**
  - "麦乐鸡" (màilè jī) is a brand name for McDonald's Chicken McNuggets. The translation "Wheat chicken" is incorrect and does not convey the intended meaning.
  - "麦乐鸡酱(màilè jī jiàng) is the brand name for McDonald's McNugget sauce. The translation "maillot chicken sauce" is also incorrect and does not convey the intended meaning.
- **Fluency**
  - "Wheat chicken" is not a natural or common English phrase. It sounds awkward and is not idiomatic.
  - "Maillot" is a French word meaning "swimsuit" or "jersey," which is completely unrelated to the source text and makes no sense in this context.
  - The phrase "no maillot chicken sauce" is not grammatically correct and is not a natural way to express the idea in English.
- **Completeness**
  - The translation fails to accurately convey the meaning of both "麦乐鸡" and "麦乐鸡酱." It does not preserve the intended meaning of the source text.

Given the above evaluation, the translation is very poor in all three criteria: accuracy, fluency, and completeness.
Score: 20 (0-100).
The translation is largely inaccurate, awkward, and fails to convey the intended meaning of the source text. A more appropriate translation would be "McNuggets, no McNugget sauce." However, since the translation provided is so far off, the score is quite low. A perfect translation would score 100, while a completely incorrect translation would score 0. Given the significant errors, a score of 20 is appropriate.

Remedy-R (14B) ref-free Evaluation Example (English-Japanese)

Source: “Which notebook is that?” Ivory asked, sitting down next to Kari on her bed.

Translation: "「哪个笔记本是你的？」アイボリがカリの横に座ってした。「きみのnotebook はどれ？」"

Remedy-R’s Evaluation:
Let’s evaluate the translation step by step based on the given criteria:

- **Accuracy: The translation attempts to convey the meaning of the source text, but there are some issues.**
  - "Which notebook is that?" is translated as "哪个笔记本是你的？" which is in Chinese, not Japanese.
  - "Ivory asked" is correctly translated as "アイボリがした".
  - "sitting down next to Kari" is translated as "カリの横に座って", which is accurate.
  - "on her bed" is missing in the translation.
- **Fluency**
  - The translation is mostly fluent, but the inclusion of "notebook" in English ("notebook") is awkward and unnatural in Japanese.
  - The phrase "哪个笔记本是你的？" is not in Japanese, which makes the translation less fluent.
- **Completeness**
  - The translation is incomplete as it does not convey "on her bed".
  - The inclusion of "notebook" in English is unnecessary and adds confusion.
- **Language Correctness**
  - The translation mixes Chinese and Japanese, which is incorrect.
  - "notebook" in English is not appropriate in a Japanese translation.

Given these points, the translation is not very good. Let’s assign a score:
- Accuracy: 50 (partially accurate but with significant errors)
- Fluency: 50 (awkward and unnatural due to language mix)
- Completeness: 50 (missing "on her bed" and unnecessary "notebook" in English)
- Language Correctness: 30 (mixing Chinese and Japanese)
Score: 45 (0-100).

Remedy-R (14B) ref-free Evaluation Example (English-German)

Source: “And with hooker Dan Frost sin-binned on the stroke of half-time, their 13 men were on the ropes and Munster reacted by engineering tries for debutant Patrick Campbell and Andrew Conway.

Translation: Und mit der Hure Dan Frost, die auf den Schlag der Halbzeit sündigte, waren ihre 13 Männer in den Seilen und Munster reagierte mit technischen Versuchen für den Debütanten Patrick Campbell und Andrew Conway.

Remedy-R’s Evaluation:
Let’s evaluate the translation step by step according to the given criteria:

- **Accuracy: The translation attempts to convey the meaning of the source text, but there are some issues.**
  - "hooker" in rugby refers to a position, not a prostitute ("Hure"). The correct term would be "Flanker" or "Ruckspitz" in German
  - "sin-binned" is a rugby term meaning "sent off for a temporary suspension." The correct translation would be "gesperrt" or "gesperrt für die Zeitung."
  - "on the ropes" is an idiomatic expression meaning "in a difficult situation." A more appropriate translation would be "in Schwierigkeiten" or "am Ende ihrer Kräfte."
  - "engineering tries" is a metaphorical way of saying "scoring tries." The correct translation would be "Tore für" or "Versuche für."
- **Fluency: The translation is mostly fluent but contains several awkward phrases and incorrect vocabulary choices.**
  - "sündigte" is not the correct translation for "sin-binned."
  - "in den Seilen" is an incorrect translation for "on the ropes."
  - "technische Versuche" is not a standard rugby term and is confusing.
- **Completeness: The translation does not accurately convey the meaning of the source text.**
  - "sin-binned" is mistranslated, which changes the meaning significantly.
  - "on the ropesïs mistranslated, which also changes the meaning.
  - "engineering tries" is mistranslated, which is confusing.

Overall, the translation is quite poor and does not meet the criteria for a high-quality translation.
Score: 50 (0-100).
The translation has significant issues with accuracy and fluency, and it does not fully convey the meaning of the source text. A score of 50 reflects these major shortcomings.

Remedy-R (14B) ref-free Evaluation Example (English-Czech)

Source: So either there's something wrong with this panel or my multimeter, or they changed the listing.

Translation: Takže bud' je něco špatně s tímto panelом, nebo s mým virometrem, nebo změnili popis.

Remedy-R's Evaluation:
Let's evaluate the translation step by step according to the given criteria:

- **Accuracy: The translation attempts to convey the meaning of the source text, but there are some inaccuracies.**
  - "panel" is correctly translated as "panel".
  - "multimeter" is incorrectly translated as "virometrem" (which means "viro-meter" in Czech). The correct translation should be "multimetro".
  - "listingïs translated as "popis", which is a reasonable approximation of "listing" in this context, but "označení" or "popis produktu" might be more precise.
- **Fluency**
  - The translation is mostly fluent and reads naturally, but "virometrem" is a non-standard term and makes the sentence sound awkward.
  - The sentence structure is mostly correct, but "virometrem" disrupts the flow.
- **Completeness**
  - The translation attempts to convey all the information from the source, but the use of "virometrem" instead of "multimetro" is a significant inaccuracy.
  - The translation is mostly complete but loses some precision due to the incorrect term.
- **Language Correctness**
  - The translation is mostly correct in Czech, but "virometrem" is not a standard term and is incorrect.
  - The use of "panelом" (with a Cyrillic "о") is a typographical error. It should be "panel".

Given these points, let's assign a score:
- **Accuracy:** 75 (due to the incorrect translation of "multimeter")
- **Fluency:** 85 (mostly fluent but disrupted by "virometrem")
- **Completeness:** 80 (mostly complete but loses precision)
- **Language Correctness:** 85 (mostly correct but "virometrem" is incorrect)
Score: 78 (0-100).

Remedy-R (14B) ref-free Evaluation Example (English-Spanish)

Source: According to the office of the commissioner for the protection of the state language, 3,692 language law violations were reported in 2023. Kiev, Kharkov, and Odessa remain "leaders in terms of the number of complaints of violations. Therefore, strict controls over the observance of the language rights of citizens in all spheres of public life should be among the priorities of community leaders this year," Taras Kremen, the country's Language Ombudsman said.

Translation: Según la oficina del comisionado para la protección del idioma oficial del estado, se reportaron 3.692 violaciones de la ley lingüística en 2023. Kiev, Kharkov y Odessa continúan siendo los "líderes" en términos del número de quejas sobre violaciones. Por lo tanto, el control estricto sobre el cumplimiento de los derechos lingüísticos de los ciudadanos en todas las esferas de la vida pública debe ser una prioridad para los líderes comunitarios este año, dijo Taras Kremen, el Defensor del Lenguaje del país.

Remedy-R's Evaluation:
Let's evaluate the translation step by step based on the given criteria:

- **Accuracy:**
  - The translation accurately conveys the meaning of the source text, including the number of reported violations, the cities mentioned, and the statement by Taras Kremen.
  - The phrase "leaders in terms of the number of complaints" is translated as "líderes en términos del número de quejas," which is a very accurate translation.
- **Fluency**
  - The translation is mostly fluent and reads naturally.
  - The sentence structure is clear and follows Spanish grammar rules.
  - There are no awkward phrasings or unnatural constructions.
- **Completeness**
  - All the information from the source text is conveyed accurately.
  - There are no unnecessary additions or omissions.
- **Language Correctness**
  - The translation is entirely in Spanish without any mixing of languages.
  - The use of "Defensor del Lenguaje" for "Language Ombudsman" is a good choice, as it is a more formal and accurate term in Spanish.

Overall, the translation is very good, with only minor room for improvement in fluency and accuracy.
Score: 95 (0-100).