# ReMedy: Learning Machine Translation Evaluation from Human Preferences with Reward Modeling

**Shaomu Tan** **Christof Monz**
Language Technology Lab
University of Amsterdam
{s.tan, c.monz}@uva.nl

## Abstract

A key challenge in MT evaluation is the inherent noise and inconsistency of human ratings. Regression-based neural metrics struggle with this noise, while prompting LLMs shows promise at system-level evaluation but performs poorly at segment level. In this work, we propose ReMedy, a novel MT metric framework that reformulates translation evaluation as a reward modeling task. Instead of regressing on imperfect human ratings directly, ReMedy learns relative translation quality using pairwise preference data, resulting in a more reliable evaluation. In extensive experiments across WMT22-24 shared tasks (39 language pairs, 111 MT systems), ReMedy achieves state-of-the-art performance at both segment- and system-level evaluation. Specifically, ReMedy-9B surpasses larger WMT winners and massive closed LLMs such as MetricX-13B, XCOMET-Ensemble, GEMBA-GPT-4, PaLM-540B, and finetuned PaLM2. Further analyses demonstrate that ReMedy delivers superior capability in detecting translation errors and evaluating low-quality translations.[1]

## 1 Introduction

Machine Translation (MT) evaluation is crucial for benchmarking progress and guiding MT development. While string-based metrics like BLEU (Papineni et al., 2002; Post, 2018), METEOR (Banerjee and Lavie, 2005), and ChrF (Popović, 2015) have been widely used since 2002, they face persistent challenges: They poorly correlate with human judgments (Freitag et al., 2022b), struggle with reliability across diverse languages (Goyal et al., 2022), and fail to distinguish between translation systems of varying quality (Przybocki et al., 2009).

Neural metrics attempt to address these shortcomings. By leveraging pre-trained multilingual language models for regression tasks, they capture semantic equivalence beyond surface-level matching and extend language coverage (Rei et al., 2020; Sellam et al., 2020). More recently, prompting Large Language Models (LLMs) for MT scoring has also shown promise in assessing translation quality across diverse contexts (Fernandes et al., 2023; Kocmi and Federmann, 2023).

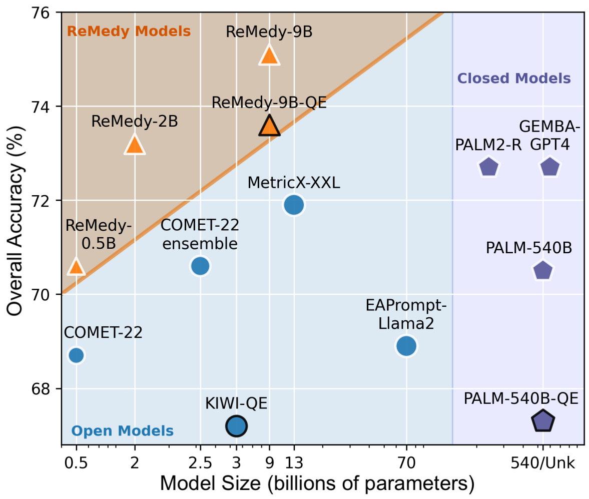

Figure 1: We report averaged accuracy over system- and segment-level pairwise accuracy for the WMT22 MQM set. The result shows that our largest ReMedy model achieves SOTA performance, surpassing previous WMT winners like MetricX-XXL, COMET, and massive closed LLMs like fine-tuned PaLM2.

However, regression-based neural metrics have limitations. Human ratings are often noisy and inconsistent due to low inter-annotator agreement (Rei et al., 2021; Song et al., 2025), making direct regression unreliable. As a result, these models tend to be less robust in real-world scenarios, particularly when detecting translation error phenomena (Amrhein et al., 2022; Moghe et al., 2025) and evaluating out-of-domain, low-quality systems compared to high-quality WMT submissions (Lo et al., 2023; Knowles et al., 2024).

Recent work (Kocmi and Federmann, 2023) also shows that prompting closed LLMs such as GPT-4

[1] https://github.com/Smu-Tan/Remedy

*Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4370–4387
November 4-9, 2025 ©2025 Association for Computational Linguistics

effectively differentiates translation quality at the system level, achieving SOTA correlations with human judgments. However, they perform substantially worse at the segment level, where individual translations are compared. This can be improved by extensive fine-tuning on MT evaluation data, yet massive LLMs like PaLM-2 still underperform much smaller models like MetricX-13B (Juraska et al., 2024). Meanwhile, small, open LLMs continue to lag behind these closed LLMs (Lu et al., 2024; Qian et al., 2024; Sindhujan et al., 2025).

In this paper, we propose **Re**ward **M**odeling for **e**valuating **d**iverse translation qualit**y** (**ReMedy**), a novel framework for MT evaluation that transforms pairwise human preferences into a robust reward signal. Unlike methods that regress over noisy absolute ratings or rely on pairwise classifiers that require quadratic comparisons, ReMedy learns from pairwise preferences, leading to more robust and reliable alignment with human judgments.

We conduct extensive experiments on the WMT22–24 metric shared tasks, spanning 39 language pairs, 111 MT systems, and about 1 million testing segments. Figure 1 shows that using the same XLM-R-large foundation, ReMedy outperforms the regression-based COMET-22 model at both segment and system levels, matching the performance of the COMET-22 ensemble (5 models). Furthermore, our ReMedy-9B model surpasses larger models such as GPT-4, PaLM-540B, fine-tuned PaLM-2, and top WMT winners like xCOMET (24B ensemble) and MetricX (13B).

Analyses on the ACES (Amrhein et al., 2022; Moghe et al., 2025) and MSLC (Lo et al., 2023) challenge sets show that ReMedy is more robust in detecting translation errors across 146 diverse language pairs and evaluating low-quality translations, making it applicable to real-world MT deployment. Beyond standard evaluation tasks, we also explore how ReMedy can be integrated into Reinforcement Learning from Human Feedback (RLHF) pipelines, leveraging its robust preference-based framework to guide model updates for improved translation quality. Our key contributions are:

**Reward Modeling for MT Assessment.** We introduce ReMedy, the first work using reward modeling for MT evaluation to achieve better alignment with human judgment than regression approaches.

**SOTA Performance with Fewer Parameters.** Our ReMedy-9B model achieves state-of-the-art results across WMT22–24 while requiring fewer parameters than the WMT winners (9B vs. 13B or 24B+) and massive LLMs like PaLM and GPT4.

**Enhanced Robustness in Challenging Scenarios.** ReMedy demonstrates superior performance in detecting translation error phenomena and reliably evaluates systems across a wide range of qualities.

**ReMedy in MT-RLHF.** We show that replacing xCOMET with ReMedy in RLHF pipelines yields consistent performance gains, demonstrating its efficacy as a reward model for improving MT quality.

## 2 Related Work

Developing MT evaluation frameworks that align with human preference has remained challenging.

**String-Based Metrics.** Metrics like BLEU (Papineni et al., 2002) and ChrF (Popović, 2015) rely on surface-level matching, which is computationally efficient but fails to capture semantic equivalence.

**Learning MT Evaluation via Regression.** Recent approaches like COMET (Rei et al., 2020, 2022), xCOMET (Guerreiro et al., 2024), and MetricX (Juraska et al., 2023, 2024) leverage pretrained multilingual models such as XLM-R (Conneau et al., 2020) and mT5 (Xue et al., 2021) to predict translation quality based on human-annotated assessments from WMT shared tasks. Despite improvements, these methods require large model sizes (>10B) or ensembles (>24B) for strong performance (Freitag et al., 2024), often misclassify low-quality translations (Lo et al., 2023), and exhibit limited robustness against diverse error phenomena (Amrhein et al., 2022).

**LLM as Judge for MT Evaluation.** Alternatively, recent work has explored using LLMs as direct judges for MT evaluation. Closed models such as GPT-4 and PaLM have competing system-level performance but struggle at the segment level (Kocmi and Federmann, 2023), even with extensive fine-tuning (Fernandes et al., 2023). Meanwhile, open LLMs perform much worse than closed ones (Qian et al., 2024) and present limitations in language inconsistency (Sindhujan et al., 2025) and prompt design (Lu et al., 2024).

**Pairwise Quality Assessment (QE).** Early works (Gamon et al., 2005; Sudoh et al., 2021) explored binary classification for MT assessment. More recently, MT-RANKER (Moosa et al., 2024) revisits this by directly optimizing the logistic regression objective, enhanced with synthetic data. However, these approaches function as classifiers rather than a standalone metric. As a result, they cannot evaluate individual translations, require quadratic comparisons for multiple systems, and operate solely as a Quality Assessment (QE) systems without leveraging available references.

## 3 ReMedy: Learning MT Metrics via Reward Modeling

In this section, we introduce ReMedy, a novel MT metric framework that learns from human preferences. We first formalize the MT evaluation task and revisit regression methods, then describe each component of ReMedy.

### 3.1 Task Definitions

Machine translation evaluation aims to assess the quality of translated text by assigning scores that correlate with human judgments. Formally, given a source sentence $src$, a candidate translation $mt$, and optionally a reference translation $ref^*$, an MT metric $M$ produces a quality score, as formalized in Eq 1. Here, $ref^*$ indicates that the reference is optional (i.e. reference-free when $ref^* = \emptyset$). Higher $M$ scores indicate better translation quality.

$$M(src, mt, ref^*) \rightarrow \mathbb{R} \tag{1}$$

### 3.2 Regression-based Approach

Recent neural MT metrics, such as COMET (Rei et al., 2020) and MetricX (Juraska et al., 2023), are trained to predict human quality ratings $h$ by minimizing the Mean Squared Error (MSE) loss (Eq. 2). However, human ratings suffer from inconsistencies and varying inter-annotator agreement. Prior work (Rei et al., 2021; Song et al., 2025) has shown that inter-annotator agreement on high-quality WMT MQM datasets yields low to modest correlation, typically ranging from 0.2 to 0.45 Kendall-Tau correlation.

$$\mathcal{L}_{mse} = \mathbb{E}_{(src,mt,ref^*,h)\in\mathcal{D}}[(M(\cdot) - h)^2] \tag{2}$$

These inconsistencies pose challenges for regression approaches, as models struggle to learn stable patterns from inherently noisy data. To mitigate this, some MT metrics normalize human ratings using z-score transformations (Rei et al., 2022; Guerreiro et al., 2024). However, Juraska et al. (2023, 2024) found that while z-normalization improves segment-level performance, it can degrade system-level performance, highlighting the trade-offs inherent in regression-based methods. These shortcomings motivate our preference-based approach.

### 3.3 ReMedy: Learn MT Metric with Pairwise Preference

Recent advances in AI alignment have demonstrated the effectiveness of reward modeling for capturing human preferences (Christiano et al., 2017) in areas such as helpfulness and safety (Ouyang et al., 2022; Bai et al., 2022). Inspired by these approaches, we propose ReMedy, an MT evaluation framework that learns to predict translation quality by modeling reward of pairwise human preferences rather than absolute scores.

**Model Architecture.** ReMedy builds on a pretrained multilingual language model with the LM head removed and a linear scoring head added to produce a scalar quality score (reward $r$). For encoder-only models, the [CLS] hidden state is mapped to the score head. For decoder-only models, following Ouyang et al. (2022) and Touvron et al. (2023), we use the hidden state of the final token as input to the linear head.

**Preference Learning Framework.** Given a input $x = \{src, ref^*\}$, and two candidate translations $y^+ = mt^+$ and $y^- = mt^-$, where human annotators prefer $mt^+$ over $mt^-$, our model learns a reward function $r_\theta(x, y)$ that assigns higher scores to preferred translations. The model is trained with a pairwise ranking objective, combined with a reward regularization term.

**Preference Ranking Loss.** The core of our method is a pairwise ranking loss based on the Bradley-Terry model (Bradley and Terry, 1952; Ouyang et al., 2022), which maximizes the probability of correctly ordering two translations according to human preference, as formalized in Eq. 3.

$$\mathcal{L}_{bt} = -\log\sigma\big(r_\theta(x, y^+) - r_\theta(x, y^-) - m(r)\big) \tag{3}$$

Here, the predicted reward scores for the preferred and non-preferred translations are denoted as $r_\theta(x, y^+)$ and $r_\theta(x, y^-)$, respectively. The margin $m(r) = h^+ - h^-$ enforces a minimum separation between scores proportional to the difference in human ratings, ensuring the model's predictions align with the degree of human preference. $\sigma$ is the sigmoid function.

This Bradley-Terry loss models the probability that translation $mt^+$ is preferred over $mt^-$ as a function of their reward difference, encouraging the model to assign higher rewards to better translations with sufficient separation when margin $m(r)$ is integrated. In our experiments, we construct preference pairs using translations $mt^+$ and $mt^-$ with their raw human ratings $h^+$ and $h^-$, given the same source and reference input.

**Reward Regularization.** We found that directly optimizing the ranking loss for MT evaluation leads to reward explosion, where the model continuously increases scalar reward scores. This occurs because the ranking loss focuses on relative differences, allowing the model to grow rewards without bound. In addition, unlike helpfulness or safety reward modeling tasks (Ouyang et al., 2022), where outputs often have large differences, translations typically differ only slightly, e.g., minor errors like omission or punctuation, and such small variations can cause the model to magnify reward discrepancies uncontrollably.

$$\mathcal{L}_{\mathit{reg}} = \mathbb{E}_r[\max(r - \beta_{\text{upper}}, 0)^2 + \max(\beta_{\text{lower}} - r, 0)^2] \tag{4}$$

To stabilize training and ensure the reward function produces well-calibrated scores within a reasonable range, we apply a reward regularization term (Eq 4). We set $\beta_{\text{upper}} = 3$ and $\beta_{\text{lower}} = -3$, to penalize rewards that exceed 3 or fall below $-3$, constraining outputs to an effective range that captures approximately 90% of the sigmoid's variation. In Section 5.2, we show that such regularization is crucial for preventing reward explosion during training, preventing degenerate performance where the model might inflate reward differences arbitrarily to satisfy the ranking objective.

**Combined Objective.** Our final training objective combines the ranking and regularization losses (Eq 5), where $\lambda$ is a hyperparameter that controls the strength of regularization. We empirically set $\lambda$ to 0.1, as higher values limit the ranking objective.

$$\mathcal{L}_{\mathit{final}} = \mathcal{L}_{\mathit{bt}} + \lambda \cdot \mathcal{L}_{\mathit{reg}} \tag{5}$$

**Inference and Reward Calibration.** While ReMedy is trained with pairwise data, it can evaluate individual triplets $(\mathit{src}, mt, \mathit{ref}^*)$ during inference to produce a scalar reward $r \in \mathbb{R}$. This avoids the quadratic comparisons of methods like MT-RANKER (Moosa et al., 2024).

Despite regularization during training, reward scores may exceed the bounds during inference in practice. To normalize rewards into the [0, 1] range and prevent clustering (which obscures quality differences), we calibrate $r$ using an entropy-guided sigmoid function $\sigma(r/\tau)$ for each language pair. The key idea is to find the optimal temperature $\tau$ by maximizing the Shannon entropy across 20 bins, encouraging an even score spread in [0, 1]. Intuitively, this prevents scores from clustering in small regions, e.g., all good translations having scores very close to 1.0.

## 4 Experimental Setup

This section outlines our benchmark choices, baselines, and implementation details.

### 4.1 Datasets and Benchmarks

We selected three complementary benchmarks to evaluate MT quality from multiple perspectives.

**WMT Metric Shared Tasks.** We use WMT22-24, standardized frameworks for comparing MT metrics. Following standard practice, we train on earlier data (from WMT17), validate on previous years, and test on the current year (See Appendix A.1). We use both high-quality Multidimensional Quality Metric (MQM) data and crowd-sourced ratings like Direct Assessment (DA)(Bojar et al., 2017) and Scalar Quality Metrics (SQM)(Mathur et al., 2020).

For evaluation, we use: WMT22 (Freitag et al., 2022a) (16 language pairs, 40 systems, 392,647 segments); WMT23 (Freitag et al., 2023) (11 language pairs, 29 systems, 282,926 segments); and WMT24 (Freitag et al., 2024) (MQM: 3 language pairs, 32 systems, 68,502 segments; ESA: 9 language pairs, 40 systems, 232,289 segments).

**ACES.** Translation Accuracy ChallengE Set (ACES) (Amrhein et al., 2022; Moghe et al., 2025) covers 146 language pairs with 68 translation error phenomena grouped into 10 types. We use ACES to analyze a wide range of translation errors, from simple perturbations to complex discourse issues.

**MSLC.** The Metric Score Landscape Challenge (MSLC) (Lo et al., 2023) evaluates metrics on low and medium-quality translations in out-of-domain contexts, using transformer MT model checkpoints

| Type | Methods | $\theta$ | ref? | System-Level | Segment-Level $acc^*_{eq}$ | | | | Avg |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Acc (3 LPs) | Avg | En-De | En-Ru | Zh-En | Corr |
| **Closed Models** | GEMBA-GPT4 (*P*) | – | ✓ | 89.8 | 55.6 | 58.2 | 55.0 | 53.4 | 72.7 |
| | PaLM (*P*) | 540B | ✓ | 90.1 | 50.8 | 55.4 | 48.6 | 48.5 | 70.5 |
| | PaLM-2 BISON (*R*) | – | ✓ | 88.0 | 57.3 | 61.0 | 51.5 | 59.5 | 72.7 |
| | PaLM-2 BISON (*GC*) | – | ✓ | 86.1 | 54.8 | 59.2 | 49.3 | 56.0 | 70.5 |
| | PaLM-2 UNICORN (*R*) | – | ✓ | 87.6 | 58.0 | **61.1** | 52.6 | 60.4 | 72.8 |
| | PaLM (*P*) | 540B | ✗ | 84.3 | 50.3 | 56.1 | 43.1 | 51.8 | 67.3 |
| | PaLM-2 BISON (*R*) | – | ✗ | 87.6 | 57.5 | $\underline{\underline{59.9}}$ | 53.4 | $\underline{\underline{59.2}}$ | 72.6 |
| **Open Models** | Skywork-RM | 8B | ✓ | 80.7 | 49.8 | 54.9 | 51.5 | 43.0 | 65.2 |
| | Llama2-EAPrompt (*P*) | 70B | ✓ | 85.4 | 52.3 | 55.2 | 51.4 | 50.2 | 68.9 |
| | COMET-22-DA (*R*) | 0.5B | ✓ | 82.8 | 54.5 | 58.2 | 49.5 | 55.7 | 68.7 |
| | COMET-22 (*R*) | 5 x 0.5B | ✓ | 83.9 | 57.3 | 60.2 | 54.1 | 57.7 | 70.6 |
| | MetricX-XXL (*R*) | 13B | ✓ | 85.0 | 58.8 | **61.1** | 54.6 | **60.6** | 71.9 |
| | COMETKIWI (*R*) | 5 x 0.5B | ✗ | 78.8 | 55.5 | 58.3 | 51.6 | 56.5 | 67.2 |
| **Ours** | ReMedy$_{\text{xlmr-22}}$ | 0.5B | ✓ | 85.8 | 55.4 | 58.3 | 52.2 | 55.8 | 70.6 |
| | ReMedy$_{\text{2B-22}}$ | 2B | ✓ | 90.5 | 55.9 | 58.0 | 53.0 | 56.6 | 73.2 |
| | ReMedy$_{\text{9B-22}}$ | 9B | ✓ | **91.2** | **58.9** | 61.0 | **60.4** | 55.4 | **75.1** |
| | ReMedy$_{\text{9B-22-QE}}$ | 9B | ✗ | $\underline{\underline{89.4}}$ | $\underline{\underline{57.8}}$ | 59.4 | $\underline{\underline{59.9}}$ | 54.2 | $\underline{\underline{73.6}}$ |

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). *P* denotes prompting; *R* and *GC* represent training with regression and generative classification objectives. **Bold** and underline indicate the best metric and QE (no reference) models.

from various training stages to create a quality spectrum. We use MSLC to assess metrics' ability to distinguish low-to-medium quality translations.

## 4.2 Baselines

We compare ReMedy with strong closed LLMs and open WMT metric winners, covering both open metrics and commercial LLM approaches.

### 4.2.1 Closed Models

**GEMBA (*P*).** A zero-shot prompting (*P*) approach using GPT-4 (Achiam et al., 2023) for quality assessment (Kocmi and Federmann, 2023).

**PaLM (*P*).** Like GEMBA, Fernandes et al. (2023) prompts PaLM-540B model (Chowdhery et al., 2023) to generate translation quality scores.

**PaLM-2 Models.** Fernandes et al. (2023) also fine-tuned PaLM-2 models using both Regression (*R*) and Generative Classification (*GC*) objectives with previous WMT data. They included BISON and UNICORN (second largest and largest in the PaLM-2 family, respectively) in the experoments.

### 4.2.2 Open Models

**Skywork-RM.** We also evaluate whether general-purpose reward models that were trained on diverse preference data (including chat, mathematics, coding, and safety) can transfer to MT evaluation. In particular, we consider Skywork-RM (Llama-3.1-8B)[2](Liu et al., 2024), the strongest open-source RM among models with ≤10B parameters on RewardBench (Lambert et al., 2025)[3]. The 8B variant is trained on Skywork-Reward-Preference-80K-v0.1, a carefully curated, high-quality preference dataset covering multiple domains. Although not MT-specific, we found at least 1000 instances containing translate or translation, suggesting some exposure to translation-related supervision.

**Llama2-EAPrompt (*P*).** The Error Analysis Prompting (Lu et al., 2024) combines chain-of-thought reasoning with error analysis to score translations, emulating human evaluation. We report their strongest open model based on Llama2-70B.

**MetricX (*R*).** A series of SOTA regression-based metrics from Google (Juraska et al., 2023, 2024), fine-tuned from mT5 models with two-stage fine-tuning and hybrid training recipes, augmented by synthetic data. We compare against the strongest MetricX variants for each year of the WMT sets.

[2] https://huggingface.co/Skywork/Skywork-Reward-Llama-3.1-8B

[3] Status as of 1 July 2025.

| Method | $\theta$ | ref? | System-Level Acc | | Segment-Level $acc^*_{eq}$ | | Avg |
|---|---|---|---|---|---|---|---|
| | | | MQM (3LPs) | SQM (8LPs) | MQM (3LPs) | SQM (8LPs) | Corr |
| Skywork-RM | 8B | ✓ | 92.4 | 89.2 | 53.6 | 56.6 | 73.0 |
| MetricX-23 (*R*) | 13B | ✓ | 90.7 | 86.3 | 56.9 | 57.0 | 72.7 |
| XCOMET-XXL (*R*) | ensemble | ✓ | 92.8 | 87.0 | 57.7 | 56.8 | 73.6 |
| GEMBA-GPT4 (*P*) | - | ✗ | **94.5** | 89.9 | 55.2 | 38.0 | 69.4 |
| MetricX-23-QE (*R*) | 13B | ✗ | 89.0 | 87.0 | 56.1 | 56.7 | 72.2 |
| XCOMET-QE (*R*) | ensemble | ✗ | 91.6 | 87.1 | 55.8 | 55.2 | 72.4 |
| COMETKIWI-XXL (*R*) | ensemble | ✗ | 91.1 | 88.7 | 54.6 | 56.0 | 72.6 |
| ReMedy$_{\text{9B-23}}$ | 9B | ✓ | 94.1 | **91.7** | **58.2** | **57.8** | **75.5** |
| ReMedy$_{\text{9B-23-QE}}$ | 9B | ✗ | 92.0 | 91.7 | 57.0 | 56.8 | 74.4 |

Table 2: Evaluation on WMT23 Metric Shared task including MQM and DA+SQM (use SQM in table for simplicity) sets. Both XCOMET-XXL and COMETKIWI-XXL are identical ensembles of 2×10.7B and 1×3.5B models.

**COMET (*R*).** COMET methods utilize XLM-R pretrained encoders to model translation quality via sentence embeddings from the source, translation, and reference. For WMT22, we compare with COMET-22-DA (0.5B) and COMET-22-ensemble (Rei et al., 2022); for WMT23-24, we compare with XCOMET (Guerreiro et al., 2024) (ensemble with 2 × 10.7B and 1 × 3.5B models).

### 4.3 Implementation and Meta-Evaluation

We train ReMedy by fine-tuning two multilingual pre-trained foundation models: XLM-R (Conneau et al., 2020) and Gemma2 (Team et al., 2024), covering both encoder- and decoder-only types. We use XLM-R-Large (0.5B) and Gemma2 2B and 9B.

We train our models using DeepSpeed (Rajbhandari et al., 2020) in bf16 precision for 1 epoch with early stopping on the validation set. We set the maximum sequence length to 1024 tokens and use the Adam optimizer with a learning rate of 5e-6 and an effective batch size of 2048. We conduct experiments using 4 NVIDIA H100 GPUs, with VLLM (Kwon et al., 2023) for fast inference.

For meta evaluation, we adopt the official WMT Metric Share Task Toolkit.[4] Following official setups, we report the Pairwise Accuracy (Acc) proposed by Kocmi et al. (2021) at system-level results for WMT22-23, and Soft Pairwise Accuracy (SPA) (Thompson et al., 2024; Freitag et al., 2024) for WMT24. For segment-level, we report pairwise accuracy with tie calibration ($acc^*_{eq}$) (Deutsch et al., 2023) for all WMT22-24, with the Perm-Both statistical significance test (Deutsch et al., 2021).

## 5 Results and Analyses

In this section, we analyze ReMedy's performance in correlating with human judgments. Our experiments show that ReMedy achieves SOTA results across WMT22-24 while maintaining parameter efficiency (Sec. 5.1). Analyses on ACES and MSLC confirm that ReMedy reliably captures diverse translation errors and quality levels (Sec. 5.3). We also show that using ReMedy in RLHF pipelines leads to consistent performance gains (Sec. 5.4).

### 5.1 Correlation with Human Preference

We evaluate ReMedy on WMT22, WMT23, and WMT24, with detailed results provided in Tables 1, 3, and 2 (see Appendix A.1 for additional details).

#### 5.1.1 ReMedy vs. Regression.

Table 1 shows that when fine-tuning the same XLM-R-Large (0.5B) foundation model, ReMedy outperforms the regression-based COMET-22-DA model by +2.6 points in system-level Acc and +0.9 in segment-level $acc^*_{eq}$ (verified by the Perm-Both statistical test). These results suggest that ReMedy delivers a more robust training signal than regression on noisy absolute ratings.

#### 5.1.2 ReMedy achieves SOTA results in WMT22–24

**WMT22**: Table 1 shows that while closed LLMs (e.g., PaLM-2) achieve high system-level accuracies, they often underperform open metrics at the segment level. Notably, ReMedy-0.5B reaches the overall performance of PaLM 540B with only 0.09% parameters. Compared to the strongest fine-tuned PaLM-2 UNICORN (*R*), ReMedy-2B exhibits a −2.1% drop in segment-level $acc^*_{eq}$, yet it

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

still presents +0.4% overall gain. Lastly, ReMedy-9B surpasses others across both system and segment levels, outperforming the strongest PaLM-2 UNICORN (*R*) by +2.3 averaged score.

**WMT23**: As presented in Table 2, ReMedy-9B outperforms winner models (XCOMET and MetricX-23) on all MQM and DA+SQM subsets including segment and system levels, with an average improvement of +1.9% and +2.8%. Furthermore, ReMedy-9B achieves these gains with significantly fewer parameters compared to the 13B MetricX-23 and ensemble XCOMET (totaling over 24B).

**WMT24**: Table 3 shows that ReMedy-9B achieves the highest rank and overall accuracy, outperforming all other methods. Furthermore, ReMedy-9B-QE outperforms all metric methods, including reference-based and -free WMT winners.

| **Methods** | **Rank** | **Avg corr** | **Sys SPA** | **Seg $acc^*_{eq}$** |
|---|---|---|---|---|
| $\text{ReMedy}_{\text{9B-24}}$ | 1 | **72.9** | 85.9 | **60.0** |
| $\text{ReMedy}_{\text{9B-24-QE}}$ | 2 | 72.1 | 84.9 | 59.3 |
| MetricX-24-Hybrid (*R*) | 3 | 72.1 | 85.6 | 58.5 |
| XCOMET-XXL (*R*) | 4 | 71.9 | **86.1** | 57.6 |
| MetricX-24-Hybrid-QE (*R*) | 5 | 71.4 | 84.9 | 58.0 |
| GEMBA-ESA (*P*) | 6 | 71.1 | 84.6 | 57.6 |
| XCOMET-XXL-QE (*R*) | 7 | 69.5 | 83.3 | 55.7 |
| Skywork-RM | 8 | 69.0 | 83.2 | 54.7 |

Table 3: Evaluation on WMT24 MQM set. We report the official accuracy percentage (SPA and $acc^*_{eq}$).

**Reference-Free ReMedy.** Although ReMedy is trained with references, the reference-free ReMedy-QE achieves SOTA performance among all QE models in WMT22-24 such as COMET-KIWI (Rei et al., 2023). Here, for the QE mode, the only difference is the reference sentence is empty, which enables multiple modes for a single model.

## 5.2 Ablation Studies

Table 4 presents the ablation studies of ReMedy, using Gemma2-2B as the foundation model.

**Reward Explosion.** We first train vanilla ReMedy, a variant optimized solely with the Bradley-Terry loss, similar to most reward models (Touvron et al., 2023; Ouyang et al., 2022). During training, we observed that the model continuously increased the final scalar reward scores regardless of the input. This behavior is intuitive, as the Bradley-Terry loss optimizes only the reward differences. In this setup, the model learns that increasing all reward scores makes the sigmoid output larger, thereby reducing the training loss. As a result, it produces excessively high rewards (mean = 17.18, std = 5.37).

Adding Reward Regularization (+ reg.) effectively mitigates this reward explosion issue, stabilizing the reward distribution (mean = 1.33, std = 0.5) and improving average accuracy on the WMT22 MQM set by +7.0%.

| **Method** | **MQM-22** | | | **Reward Dist** | |
|---|---|---|---|---|---|
| | **Sys** | **Seg** | **Avg** | **Mean** | **Std** |
| Vanilla-ReMedy-2B | 79.6% | 52.2% | 65.9% | 17.18 | 5.37 |
| + reg. | 90.9% | 54.9% | 72.9% | 1.33 | 0.50 |
| + reg. + margin | 89.8% | 55.2% | 72.5% | 1.93 | 0.63 |
| + reg. + margin + cali. | 90.5% | 55.9% | 73.2% | 0.82 | 0.08 |

Table 4: Performance and reward distribution of adding reward regularization (reg.), margin, and reward calibration (cali.) for ReMedy-2B on WMT22 test set.

**Margin and Inference Calibrations.** Incorporating the rating difference as a margin signal enhances segment-level performance (+0.3 Acc) by informing the model about the degree of preference between translations. For reward calibration, we apply a sigmoid function with its temperature guided by entropy (see Section 3.3). This calibration normalizes rewards to the [0,1] range while preserving meaningful distinctions between translations of similar quality, slightly improving overall performance by +0.7%. Notably, ReMedy achieves SOTA performance without calibration, which only serves for normalization purposes. Note that calibration only improves tie situations, see our detailed analyses in Appendix A.4

## 5.3 Analyses on Challenge sets

In addition to the WMT benchmarks, we analyze ReMedy's performance in detecting translation errors and out-of-domain low-quality translations.

**MSLC Challenge Set.** On the MSLC challenge set, ReMedy provides reliable quality scores across a wide range of translation outputs, effectively distinguishing between low- and medium-quality translations. As shown in Figure 2, unlike XCOMET and MetricX, ReMedy presents a clear quality boundary for the English-German MT model for its different checkpoints, especially for out-of-domain low and medium quality (corresponding to 1 to 16 BLEU scores) translations.

| | ref? | Add | Omi | Mis-T | Un-T | DNT | Over | Under | RW-K | WL | Punc | ACES |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BLEU | ✓ | 0.75 | 0.44 | -0.23 | 0.36 | 0.60 | -0.84 | -0.86 | -0.77 | 0.66 | 0.64 | -2.8 |
| ChrF | ✓ | 0.64 | 0.78 | 0.16 | **0.78** | **0.96** | -0.70 | -0.59 | -0.29 | **0.69** | 0.74 | 3.71 |
| MetricX-13B | ✓ | -0.10 | 0.53 | 0.58 | 0.65 | 0.88 | 0.75 | 0.55 | 0.71 | -0.32 | 0.37 | 13.54 |
| COMET-22 | ✓ | 0.33 | 0.81 | 0.57 | 0.54 | 0.90 | 0.69 | 0.54 | 0.57 | -0.32 | 0.54 | 16.41 |
| KG-BERTScore | ✓ | **0.79** | 0.81 | 0.49 | -0.46 | 0.76 | 0.65 | 0.53 | 0.49 | 0.31 | 0.26 | 17.49 |
| COMET-KIWI-22 | ✗ | 0.36 | 0.83 | 0.63 | 0.23 | 0.78 | 0.74 | 0.57 | 0.58 | -0.36 | 0.49 | 16.95 |
| MT-Ranker-13B | ✗ | 0.65 | **0.97** | 0.63 | 0.25 | 0.84 | 0.63 | 0.54 | 0.66 | -0.53 | **0.97** | 18.46 |
| ReMedy$_{\text{2B-22}}$ | ✓ | 0.35 | 0.72 | 0.66 | 0.63 | 0.70 | 0.79 | 0.55 | 0.82 | 0.20 | 0.64 | 17.74 |
| ReMedy$_{\text{9B-22}}$ | ✓ | 0.49 | 0.86 | 0.71 | 0.70 | 0.76 | **0.81** | 0.56 | **0.89** | 0.31 | 0.60 | **19.90** |
| ReMedy$_{\text{2B-22-QE}}$ | ✗ | 0.05 | 0.69 | 0.67 | 0.11 | 0.50 | 0.73 | 0.52 | 0.76 | -0.17 | 0.52 | 14.49 |
| ReMedy$_{\text{9B-22-QE}}$ | ✗ | 0.48 | 0.81 | **0.73** | 0.39 | 0.56 | **0.81** | **0.59** | 0.87 | 0.04 | 0.58 | 18.93 |

Table 5: Kendalls tau-like correlation results for the ten error categories spaning 68 translation phenomena for 146 language pairs. ACES-Score represents the overall performance across all categories (see A.1.2). Addition (Add), Omission (Omi), Mis-T (Mistranslation), Un-T (Untranslated), DNT (Do Not Translate), Over (Overtranslation), Under (Undertranslation), RW-K (Real-World Knowledge), WL (Wrong Language), Punc (Punctuation).

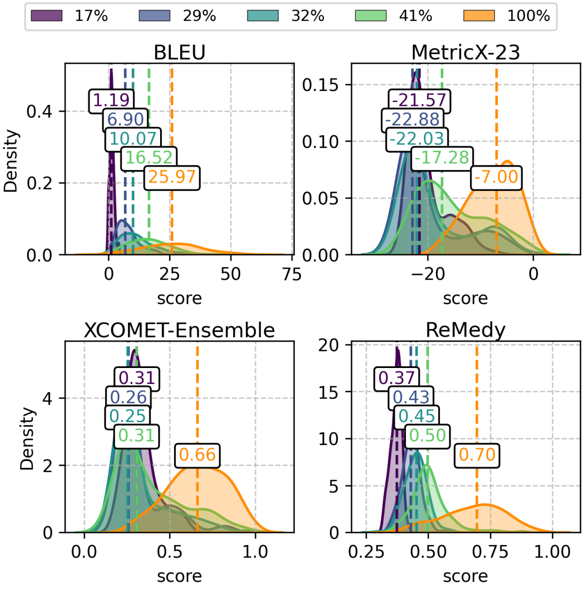

Figure 2: Kernel density plots of quality scores at various model checkpoints. Percentages indicate training progress stages, with dashed lines marking mean scores.

**ACES Challenge Set.** As shown in Table 5, ReMedy-9B achieves new SOTA results on the ACES benchmark that covers 146 language pairs, demonstrating the highest overall correlation (ACES score) with human judgments in detecting 68 diverse translation error phenomena.

We noticed that all neural metrics perform poorly on the Wrong Language (WL) phenomenon. This is intuitive since such errors contain semantically equivalent but off-target (Tan and Monz, 2023) translations. Incorporating synthetic data holds promise, but we leave this to future work.

### 5.4 ReMedy in RLHF Pipelines

Lastly, we integrated ReMedy as a reward model in Reinforcement Learning from Human Feedback (RLHF) pipelines. We implement Contrastive Preference Optimization (CPO) (Xu et al., 2024b) based on the ALMA-13B (Xu et al., 2024a) model. Following the original CPO setup, we keep the training data unchanged, then use ReMedy-9B to score References, GPT-4 and ALMA translations. We conduct CPO tuning on ALMA-13B with LoRA using the same hyper-parameter, then evaluate the final models with greedy decoding.

| RM | BLEU | COMET22 | KIWI | XCOMET | ReMedy |
|---|---|---|---|---|---|
| Results on WMT22 Testset (10 LPs) | | | | | |
| XCOMET | 28.6 | 85.6% | 81.9% | 90.2% | 80.8% |
| ReMedy | **29.8** | **85.9%** | **82.3%** | **90.3%** | **81.1%** |
| Results on WMT23 Testset (6 LPs) | | | | | |
| XCOMET | 28.0 | 83.0% | 76.9% | 88.1% | 80.6% |
| ReMedy | **29.4** | **83.3%** | **77.1%** | **88.2%** | **81.1%** |

Table 6: Performance of using XCOMET and ReMedy-9B as reward models for ALMA13B-CPO tuning on WMT22 and WMT23 general MT testsets.

To avoid metric interference (Pombal et al., 2025), i.e., use the same metrics for both model tuning and evaluation, we report results on various metrics including BLEU, KIWI-10B, XCOMET-10.9B, and ReMedy-9B. Table 6 shows that replacing the XCOMET reward model with ReMedy-9B yields consistent performance gains on all metric scores, underscoring ReMedy's versatility and potential for downstream MT improvements.

## 6 Conclusions

To address the challenges of noisy and inconsistent human ratings in MT evaluation, we introduced ReMedy, a novel framework leveraging reward modeling, augmented by reward regularization and calibration, to learn directly from pairwise human preferences. Our extensive experiments on WMT22–24 demonstrate that ReMedy achieves state-of-the-art performance at both segment and system levels. Notably, our 9B parameter ReMedy model surpasses significantly larger models, including GPT-4, PaLM-540B, XCOMET-Ensemble, and MetricX-13B. Further analyses confirmed its robustness on challenge sets designed to test error detection and handling of varying quality levels. Additionally, ReMedy's integration into RLHF pipelines highlights its potential as an effective reward model for improving MT systems. ReMedy shows that reward modeling with preference learning offers a more robust, efficient, and human-aligned approach to machine translation evaluation.

## Limitations

In this paper, we do not include the utilization of synthetic data in MT evaluation. Previous studies such as MetricX (Juraska et al., 2024), XCOMET (Guerreiro et al., 2024) found constructing synthetic data for out-of-domain and fine-grained translation errors can improve the overall performance and form more robust systems. In this work, we focus more on how to improve the MT metric system with current available open-source data. However, ReMedy holds great promise in leveraging synthetic data, since it only requires pairwise preference data rather than absolute ratings like MetricX or XCOMET requires for regression, we leave this to future work.

We noticed that for the WMT24 ESA subset, ReMedy-9B-24 performs slightly worse than MetricX and XCOMET (see Appendix A.3). Specifically, we found gaps mostly on English-Hindi and English-Icelandic pairs, where LLM-based approaches like GEMBA-ESA also present lower performance. We hypothesize this could be due to the nature of these language pairs remaining low-resource for pre-trained decoder-only LLMs. Nonetheless, we found that ReMedy-9B-22 outperforms MetricX and COMET on unseen extremely low-resource language pairs like English-Livonian, and Yakut-Russian in the WMT22 test set. We plan to look at the potential reasons in the future.

## Broader Impact

We acknowledge several ethical considerations in MT evaluation research. To address the risk of mistranslation, we prioritize high-quality data from WMT Metric Shared tasks, though fairness challenges persist as metrics may perform inconsistently across the linguistic spectrum, particularly for low-resource languages. Furthermore, MT systems and evaluation metrics can perpetuate societal biases present in training data, such as human biases.

## Acknowledgments

This research was funded in part by the Netherlands Organization for Scientific Research (NWO) under project numbers VI.C.192.080 and 2023.017. We thank SURF (www.surf.nl) for the support in using the National Supercomputer Snellius.

## 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*.

Chantal Amrhein, Nikita Moghe, and Liane Guillou. 2022. ACES: Translation accuracy challenge sets for evaluating machine translation metrics. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 479–513, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017. Results of the WMT17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 489–513, Copenhagen, Denmark. Association for Computational Linguistics.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106, Columbus, Ohio. Association for Computational Linguistics.

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.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

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.

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. 2021a. 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, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021b. 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. 2022a. 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.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André FT Martins. 2022b. 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.

Michael Gamon, Anthony Aue, and Martine Smets. 2005. Sentence-level mt evaluation without reference translations: Beyond language modeling. In *Proceedings of the 10th EAMT Conference: Practical applications of machine translation*.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, MarcAurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

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.

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 and Christian Federmann. 2023. 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.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, Lester James Validad Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. 2025. Rewardbench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1755–1797.

Chris Yuhao Liu, Liang Zeng, Jiacai Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. 2024. Skywork-reward: Bag of tricks for reward modeling in llms. *arXiv preprint arXiv:2410.18451*.

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, 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, Online. Association for Computational Linguistics.

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.

Ibraheem Muhammad Moosa, Rui Zhang, and Wenpeng Yin. 2024. Mt-ranker: Reference-free machine translation evaluation by inter-system ranking. In *12th International Conference on Learning Representations, ICLR 2024*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

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.

José Pombal, Nuno M Guerreiro, Ricardo Rei, and André FT Martins. 2025. Adding chocolate to mint: Mitigating metric interference in machine translation. *arXiv preprint arXiv:2503.08327*.

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.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Mark Przybocki, Kay Peterson, Sébastien Bronsart, and Gregory Sanders. 2009. The nist 2008 metrics for machine translation challengeoverview, methodology, metrics, and results. *Machine Translation*, 23:71–103.

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.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE.

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, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André FT Martins, and Alon Lavie. 2021. Are references really needed? unbabel-ist 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040.

Ricardo Rei, Nuno M Guerreiro, Daan van Stigt, Marcos Treviso, Luísa Coheur, José GC de Souza, André FT Martins, et al. 2023. Scaling up cometkiwi: Unbabel-ist 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848.

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.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Archchana Sindhujan, Diptesh Kanojia, Constantin Orasan, and Shenbin Qian. 2025. When llms struggle: Reference-less translation evaluation for low-resource languages. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 437–459.

Yixiao Song, Parker Riley, Daniel Deutsch, and Markus Freitag. 2025. Enhancing human evaluation in machine translation with comparative judgment. *arXiv preprint arXiv:2502.17797*.

Katsuhito Sudoh, Kosuke Takahashi, and Satoshi Nakamura. 2021. Is this translation error critical?: Classification-based human and automatic machine translation evaluation focusing on critical errors. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 46–55.

Shaomu Tan and Christof Monz. 2023. Towards a better understanding of variations in zero-shot neural machine translation performance. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore. Association for Computational Linguistics.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

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.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

# A Appendix

## A.1 Data

### A.1.1 WMT Metric Shared Tasks

WMT Metric Shared Tasks provided a standardized framework for comparing automatic MT evaluation metrics using human assessments since 2008 (Callison-Burch et al., 2008). In WMT22-24, various annotation methods have been employed. Among these, the Multidimensional Quality Metric (MQM) stands out due to its reliance on professional translators for fine-grained error annotations, making it particularly reliable for assessing high-quality MT outputs (Freitag et al., 2021a).

In contrast, other evaluation approaches including Direct Assessment (DA) (Bojar et al., 2017), Scalar Quality Metrics (SQM) (Mathur et al., 2020), Error Span Analysis (ESA) (Freitag et al., 2024) are based on crowdsourced ratings, which may not always capture the same level of nuance and precision (Freitag et al., 2021a).

Human assessments in WMT22-24 include four types of annotations below. MQM is considered as the highest-quality assessment, which is more reliable for high-quality MT predictions (Freitag et al., 2021a).

- **Multidimensional Quality Metric (MQM)**: Professional translators provide fine-grained error annotations (Freitag et al., 2021b).
- **Direct Assessment (DA)**: Crowdsourced holistic quality ratings on a 0–100 scale (Bojar et al., 2017).
- **Scalar Quality Metrics (SQM)** (Mathur et al., 2020): A simplified version of MQM with fewer error categories.
- **Error Span Analysis (ESA)** (Freitag et al., 2024): 0–100 Ratings accompanied by error span annotations.

Following standard practice (Guerreiro et al., 2024), we train on earlier data (e.g., WMT17), validate on previous years, and test on the current year (see Table 7 for details). Our evaluations are conducted on official WMT22–24 datasets. WMT22 (Freitag et al., 2022a): Contains MQM and DA+SQM subsets with 16 language pairs, 40 systems, and 392,647 segments. WMT23 (Freitag et al., 2023): Includes 282,926 segments over 11 language pairs and 29 MT systems. WMT24 (Freitag et al., 2024): For the high-quality MQM subset, there are 3 language pairs, 32 systems, and 68,502 segments; the ESA subset includes 232,289 segments covering 9 language pairs and 40 systems.

### A.1.2 ACES Score

In this paper, we follow the original ACES Score calculation (Amrhein et al., 2022; Moghe et al., 2025), which provides a comprehensive assessment by combining performance on various error types with appropriate weightings. As shown in Equation 6, the ACES Score assigns higher weights (5) to critical error categories such as addition, omission, mistranslation, overtranslation, and undertranslation, while giving lower weights to categories like untranslated segments (1), wrong language (1), and punctuation errors (0.1).

$$\text{ACES} = sum \left\{ \begin{array}{c} 5 * \tau_{\text{addition}} \\ 5 * \tau_{\text{omission}} \\ 5 * \tau_{\text{mistranslation}} \\ 1 * \tau_{\text{untranslated}} \\ 1 * \tau_{\text{do not translate}} \\ 5 * \tau_{\text{overtranslation}} \\ 5 * \tau_{\text{undertranslation}} \\ 1 * \tau_{\text{real-world knowledge}} \\ 1 * \tau_{\text{wrong language}} \\ 0.1 * \tau_{\text{punctuation}} \end{array} \right\} \quad (6)$$

This weighting scheme reflects the relative impact of different error types on overall translation quality. For more details on the ACES challenge set and the development of this scoring methodology, we refer readers to Amrhein et al. (2022) and Moghe et al. (2025).

## A.2 Pairwise Data Construction for Reward Modeling

We construct pairwise preference training and validation data using the original raw human ratings for each translation. Specifically, given the same source and reference sentence pair $(src, ref)$, we examine human ratings for different translations and construct preference pairs $(mt^+, mt^-)$ where the human rating for $h_{mt^+}$ is higher than that for $h_{mt^-}$.

For DA (Direct Assessment) data with a [0,100] scale, we set a rating difference threshold of 25 points, following the common understanding that translations differing by less than 25 points should be considered of equivalent quality.

For MQM (Multidimensional Quality Metrics) data with a [0,25] scale, we use a much smaller threshold of 0.1, as MQM annotations are more fine-grained, where even small differences like punctuation errors can meaningfully impact translation quality.

Once we construct the pairwise preference data, we format inputs differently depending on the foundation model architecture (see Figure 3 for more details).

Finally, we evaluate ReMedy with the official testset directly for each individual translation $(src, mt, ref^*)$, without doing any data preprocessing steps. The final meta-evaluation is done by the official MTME tool.

| Train set | Val set | Benchmark/Test set | #Languages in Test set | #Segments in Test set | Subsets in Test set |
|---|---|---|---|---|---|
| WMT17-20 | WMT21 | WMT22 | 16 language pairs | 392,647 segments | MQM, DA |
| WMT17-21 | WMT22 | WMT23 | 11 language pairs | 282,926 segments | MQM, DA+SQM |
| WMT17-22 | WMT23 | WMT24 | 12 language pairs | 232,289 segments | MQM, ESA |

Table 7: WMT22-24 Benchmark Descriptions.

## A.3 Additional Results

### A.3.1 WMT22

We list the full results of WMT22 in Table 8, demonstrating the performance of various metric systems on both MQM and DA subsets. Note that all closed models and Llama2-EAPrompt do not validate their results on the DA set.

### A.3.2 WMT24

For WMT24, we present the full results for both MQM and ESA subsets in Table 9. Following the WMT24 official meta evaluation protocol (Freitag et al., 2024), we use the MQM subset for our primary comparisons as it provides higher-quality human annotations than the crowd-sourced ESA set. Our analysis reveals that ReMedy-9B-24 performs slightly worse on the ESA subset, primarily due to lower performance on English-Hindi and English-Icelandic language pairs (complete evaluation results available in our repository[5]).

This underperformance likely stems from these languages being relatively low-resource in the pretrained Gemma2 model. Interestingly, ReMedy-9B-22 still outperforms MetricX and COMET on previously unseen extremely low-resource language pairs such as English-Livonian and Yakut-Russian in the WMT22 test set. We intend to investigate these performance differences in future work.

## A.4 Reward Calibration Analysis

In this section, we demonstrate how our entropy-guided temperature selection adapts to different reward distributions, maximizing the information content of the final calibrated scores. Note that such calibration does not change the ranking of evaluated translations, thus, it can only improves the segment pairwise accuracy for tie situations.

By selecting the temperature that maximizes Shannon entropy across 20 uniform bins in the [0,1] interval, we ensure calibrated scores utilize the full range effectively, preventing clustering and preserving meaningful distinctions between translations of varying quality. Our entropy maximization can be formulated below in Eq 7:

$$\tau^* = \arg\max_{\tau} H(P_\tau) = \arg\max_{\tau} -\sum_{i=1}^{20} p_i^\tau \log p_i^\tau \tag{7}$$

where $P_\tau = \{p_1^\tau, p_2^\tau, ..., p_{20}^\tau\}$ represents the distribution of calibrated scores across 20 bins when using temperature $\tau$. This approach dynamically adapts to different reward distributions, providing optimal discrimination where it matters most.

In our experiments, we apply the reward calibration for each language pair, since we found different language pairs could demonstrate various translation quality in general, e.g., high resource language pairs like English-German generally have higher translation quality than low-resource language pairs.

### A.4.1 High Temperature Case Study: Right-Skewed Distributions

Figure 4 illustrates our entropy-guided reward calibration for WMT22 English-German translation submissions. For high-quality MT systems, raw rewards are typically concentrated in the upper range, creating a right-skewed distribution.

The top panel shows two sigmoid functions with different temperature values: the standard sigmoid with $T = 1.0$ (blue) and our entropy-optimized sigmoid with $T = 1.8$ (red). The mathematical formulations display how the temperature parameter affects the steepness of the curve. The bottom panel shows the histogram of raw reward values from ReMedy-9B-22, where rewards are heavily concentrated between 4 and 6, reflecting the high quality of WMT22 English-German translation submissions.

With a standard sigmoid ($T = 1.0$), most high reward values would be mapped to scores very close to 1.0, making distinguishing between good and excellent translations difficult. By increasing the temperature to $T = 1.8$, the sigmoid curve

[5] https://github.com/Smu-Tan/Remedy

| Type | Methods | $\theta$ | ref? | System-Level Acc | | Segment-Level Acc | | Avg Corr | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | MQM | DA | MQM | DA | MQM | DA | All |
| **Closed Models** | GEMBA-ChatGPT (*P*) | 175B | ✓ | 81.0% | - | 50.1% | - | 65.6% | - | - |
| | GEMBA-GPT4 (*P*) | - | ✓ | 89.8% | - | 55.6% | - | 72.7% | - | - |
| | PaLM (*P*) | 540B | ✓ | 90.1% | - | 50.8% | - | 70.5% | - | - |
| | PaLM-2 BISON (*R*) | <340B | ✓ | 88.0% | - | 57.3% | - | 72.7% | - | - |
| | PaLM-2 BISON (*GC*) | <340B | ✓ | 86.1% | - | 54.8% | - | 70.5% | - | - |
| | PaLM-2 UNICORN (*R*) | ~340B | ✓ | 87.6% | - | 58.0% | - | 72.8% | - | - |
| | PaLM (*P*) | 540B | ✗ | 84.3% | - | 50.3% | - | 67.3% | - | - |
| | PaLM-2 BISON (*R*) | - | ✗ | 87.6% | - | 57.5% | - | 72.6% | - | - |
| | PaLM-2 BISON (*GC*) | - | ✗ | 86.1% | - | 53.2% | - | 60.7% | - | - |
| | PaLM-2 UNICORN (*GC*) | - | ✗ | 86.1% | - | 52.9% | - | 69.5% | - | - |
| **Open Models** | Llama2-EAPrompt (*P*) | 70B | ✓ | 85.4% | - | 52.3% | - | 68.9% | - | - |
| | COMET-22-DA (*R*) | 0.5B | ✓ | 82.8% | 86.4% | 54.5% | 55.4% | 68.7% | 70.9% | 69.8% |
| | COMET-22 (*R*) | ensemble | ✓ | 83.9% | 85.8% | 57.3% | **57.2%** | 70.6% | 71.5% | 71.0% |
| | MetricX-XXL (*R*) | 13B | ✓ | 85.0% | 86.5% | 58.8% | 55.6% | 71.9% | 71.1% | 71.5% |
| | Llama2-EAPrompt (*P*) | 70B | ✗ | 85.8% | - | 52.0% | - | 68.9% | - | - |
| | COMETKiwi (*R*) | ensemble | ✗ | 78.8% | 85.4% | 55.5% | 56.5% | 67.2% | 71.0% | 69.1% |
| | ReMedy$_{\text{xlm-r}}$ (Ours) | 0.5B | ✓ | 85.8% | 86.6% | 55.4% | 55.6% | 70.6% | 71.1% | 70.9% |
| | ReMedy$_{\text{gemma2}}$ (Ours) | 2B | ✓ | 90.5% | 86.2% | 55.9% | 53.9% | 73.2% | 70.0% | 71.6% |
| | ReMedy$_{\text{gemma2}}$(Ours) | 9B | ✓ | **91.2%** | **87.7%** | **58.9%** | 56.0% | **75.1%** | **71.9%** | **73.5%** |
| | ReMedy$_{\text{gemma2}}$ (Ours) | 9B | ✗ | 89.4% | 85.8% | 57.8% | 54.3% | 73.6% | 70.0% | 71.8% |

Table 8: Evaluation on WMT22 MQM (3 LPs) and DA (13 LPs) set. The system-level results are Pairwise Accuracy proposed by Kocmi et al. (2021), and segment-level results are based on the group-by-item pairwise accuracy with tie calibration (Deutsch et al., 2023). P denotes prompting (no tuning); R and GC represent Regression and Generative Classification training objectives. **Bold** and underline indicate the best metric and QE (no reference) models. COMET-22 and COMETKiwi are ensembled with 5x and 6x 0.5B models, respectively.

| Methods | $\theta$ | ref? | System-Level SPA | | Segment-Level $acc^*_{eq}$ | | Avg Corr | |
|---|---|---|---|---|---|---|---|---|
| | | | MQM | ESA | MQM | ESA | MQM | ESA |
| XCOMET (*R*) | 24B | ✓ | **86.1%** | 85.4% | 57.6% | 56.3% | 71.9% | 70.9% |
| MetricX-24-Hybrid (*R*) | 13B | ✓ | 85.6% | **86.3%** | 58.5% | 56.5% | 72.1% | **71.4%** |
| ReMedy$_{\text{9B-24}}$ (Ours) | 9B | ✓ | 85.9% | 84.8% | **60.0%** | 55.2% | **72.9%** | 70.0% |
| GEMBA-ESA (*P*) | - | ✗ | 84.6% | 81.5% | 57.6% | 42.2% | 71.1% | 61.8% |
| XCOMET-QE (*R*) | 24B | ✗ | 83.3% | 83.9% | 55.7% | 55.1% | 69.5% | 69.5% |
| MetricX-24-Hybrid-QE (*R*) | 13B | ✗ | 84.9% | 84.2% | 58.0% | 55.5% | 71.4% | 69.8% |
| ReMedy$_{\text{9B-24-QE}}$ (Ours) | 9B | ✗ | 84.9% | 83.5% | 59.3% | 54.2% | 72.1% | 68.9% |

Table 9: Evaluation on WMT24 MQM (3 LPs) and ESA (9 LPs) set. **Bold** and underline indicate the best metric and QE (no reference) models.

is horizontally stretched, creating more separation between high-quality translations in the final [0,1] score range. The vertical dashed red lines illustrate how specific histogram bins map to points on the sigmoid curve.

Table 10 shows numerically how the increased temperature creates meaningful separation between high-quality translations. For example, raw scores of 4.09 and 5.00 would receive nearly identical scores (0.984 vs. 0.993) with the standard sigmoid, but more distinguishable scores (0.907 vs. 0.941) with our calibrated approach.

### A.4.2 Low-Normal Temperature Case Study: Rewards with Normally Distribution

On the other hand, Figure 5 demonstrates calibration for a more evenly distributed set of raw rewards (approximately normally distributed around 0). Such distribution appears when evaluating diverse MT systems with varying quality levels. Here,

| Raw score ($x$) | $\sigma(x, T = 1.0)$ | $\sigma(x, T = 1.8)$ |
|---|---|---|
| -3.00000 | 0.04743 | 0.15887 |
| 2.25000 | 0.90465 | 0.77730 |
| 3.50000 | 0.97069 | 0.87484 |
| 4.09375 | 0.98360 | 0.90673 |
| 4.50000 | 0.98901 | 0.92414 |
| 4.75000 | 0.99142 | 0.93332 |
| 5.00000 | 0.99331 | 0.94146 |
| 5.15625 | 0.99427 | 0.94607 |
| 5.28125 | 0.99494 | 0.94950 |
| 5.40625 | 0.99553 | 0.95273 |
| 5.53125 | 0.99605 | 0.95576 |
| 5.62500 | 0.99641 | 0.95791 |
| 5.71875 | 0.99673 | 0.95996 |
| 5.87500 | 0.99720 | 0.96317 |
| 6.40625 | 0.99835 | 0.97232 |

Table 10: Sigmoid calibration values for right-skewed reward distributions of high-quality submission systems. The higher temperature ($T = 1.8$) creates larger separation between high rewards that would otherwise cluster near 1.0 with the standard sigmoid. This enables better discrimination between good and no-error translations. These scores correspond to values in Figure 4.

our entropy-guided approach selects a temperature of $T = 0.7$, lower than the standard $T = 1.0$.

With $T = 0.7$, the sigmoid curve is more compressed, making it steeper around the center. This compression provides enhanced discrimination for translations in the mid-quality range, where most reward values are concentrated in this distribution. The histogram in the bottom panel confirms the balanced distribution of raw rewards, and the vertical dashed lines illustrate the mapping between histogram bins and sigmoid values.

Table 11 demonstrates how the lower temperature creates greater separation in the central region of the distribution. For instance, raw scores of -0.76 and 0.52 show larger differences with $T = 0.7$ (0.253 vs. 0.677) compared to $T = 1.0$ (0.319 vs. 0.627), improving our ability to discriminate between average-quality translations.

These case studies demonstrate how our entropy-guided temperature selection dynamically adapts to different reward distributions. This approach is proved to yield better alignment with human judgments (see Table 4) when evaluating diverse MT systems that may produce translations clustered in different quality ranges, ensuring optimal discrimination across the entire quality spectrum.

| Raw score ($x$) | $\sigma(x, T = 1.0)$ | $\sigma(x, T = 0.7)$ |
|---|---|---|
| -3.78125 | 0.02229 | 0.00449 |
| -2.68750 | 0.06371 | 0.02106 |
| -2.25000 | 0.09535 | 0.03863 |
| -1.78906 | 0.14319 | 0.07204 |
| -1.42188 | 0.19437 | 0.11596 |
| -1.06250 | 0.25683 | 0.17978 |
| -0.75781 | 0.31912 | 0.25302 |
| -0.49219 | 0.37938 | 0.33112 |
| -0.20703 | 0.44843 | 0.42659 |
| 0.14062 | 0.53510 | 0.55005 |
| 0.51953 | 0.62704 | 0.67747 |
| 0.92188 | 0.71542 | 0.78868 |
| 1.35938 | 0.79566 | 0.87457 |
| 1.87500 | 0.86704 | 0.93575 |
| 2.96875 | 0.95114 | 0.98581 |

Table 11: Sigmoid calibration values for evenly distributed rewards. The lower temperature ($T = 0.7$) creates larger separation in the central region where most scores are concentrated, improving discrimination between translations of moderate quality. These scores correspond to values in Figure 5.

**Encoder-only models.** For encoder-only models, we use a simple concatenation format:

```
# Preferred translation pair
    chosen = [
        {src_lang}: {src}, {Reference}: {ref*}, {tgt_lang}: {mt+}
    ]
# Non-preferred translation pair
    rejected = [
        {src_lang}: {src}, {Reference}: {ref*}, {tgt_lang}: {mt-}
    ]
```

**Decoder-only models.** For decoder-only models, we use a chat template format with paired preferred and non-preferred examples:

```
# Preferred translation pair
    chosen = [
        {’role’: ’user’,
         ’content’: "Translate the following {src_lang} text into
                    natural, fluent {tgt_lang} sentence while
                    preserving the original meaning. You are also
                    given a translation template.
                    {src_lang}:{src}
                    Template:{ref*}
                    {tgt_lang}:"},
        {’role’: ’assistant’, ’content’: {mt+}}
    ]
# Non-preferred translation pair
    rejected = [
        {’role’: ’user’,
         ’content’: "Translate the following {src_lang} text into
                    natural, fluent {tgt_lang} sentence while
                    preserving the original meaning. You are also
                    given a translation template.
                    {src_lang}:{src}
                    Template:{ref*}
                    {tgt_lang}:"},
        {’role’: ’assistant’, ’content’: {mt-}}
    ]
```

Where {src_lang}, {tgt_lang} represent source and target language, $src$, $ref^*$ denote the source and reference sentences, and $mt^+$ and $mt^-$ represent the preferred and non-preferred translations.

Figure 3: ReMedy data format for training and inference

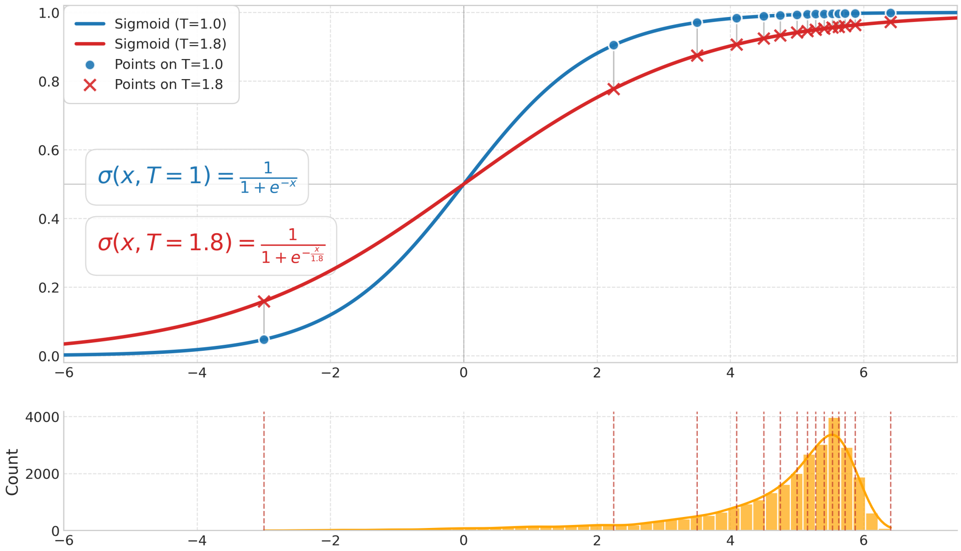

Figure 4: Reward calibration with high temperature. For such distributions, raw rewards are typically concentrated in the upper range, creating a right-skewed distribution. With a standard sigmoid ($T = 1.0$), most high reward values would be mapped to scores very close to 1.0, making distinguishing between good and excellent translations difficult. By increasing the temperature to $T = 1.8$, the sigmoid curve is horizontally stretched, creating more separation between high-quality translations in the final [0,1] score range.

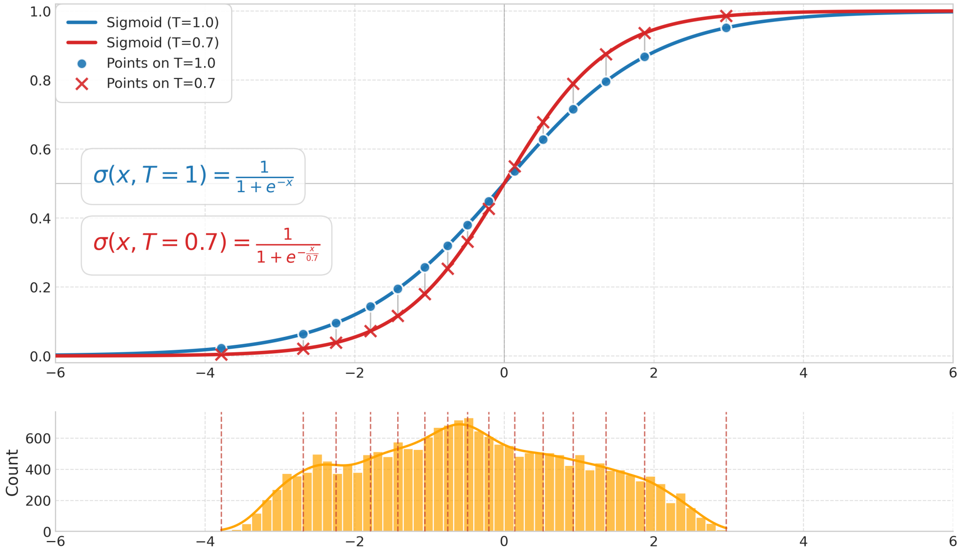

Figure 5: Reward calibration with low temperature. Such distribution appears when evaluating diverse MT systems with varying quality levels. With $T = 0.7$, the sigmoid curve is more compressed, making it steeper around the center. This compression provides enhanced discrimination for translations in the mid-quality range, where most reward values are concentrated in this distribution.