PEAR: Pairwise Evaluation for Automatic Relative scoring

Anonymous ACL submission

Abstract

We present PEAR (Pairwise Evaluation for Automatic Relative scoring), a supervised QE metric family that reframes reference-free MT evaluation as a graded pairwise comparison. Given a source segment and two candidate translations, PEAR predicts the direction and magnitude of their quality difference. PEAR learns from pairwise supervision constructed by differencing human segment-level judgments under an antisymmetry-consistent objective.

On the WMT24 meta-evaluation benchmark, PEAR outperforms strictly matched single-candidate QE baselines trained with the same data and backbones, isolating the benefit of the proposed pairwise formulation. Despite using substantially fewer parameters than recent large WMT submissions, PEAR surpasses far larger QE models and strong reference-based metrics. Inter-metric analyses further indicate that PEAR yields a less redundant evaluation signal relative to other top metrics. Finally, we show that PEAR is a strong utility for Minimum Bayes Risk decoding, and that an antisymmetry-based shortcut reduces pairwise scoring cost with negligible impact.

1 Introduction

Automatic metrics are a primary tool for comparing modern Machine Translation (MT) systems. Shared-task evaluations and much of the research literature rely heavily on metric scores (Marie et al., 2021; Kocmi et al., 2021, 2024c). Human evaluation remains the highest-quality signal but is expensive and difficult to scale across the growing number of systems, domains, and language pairs of interest (Zouhar et al., 2025a). As MT quality improves, differences between strong systems become subtle, making both human and automatic discrimination harder (Proietti et al., 2025a,b; Zouhar et al., 2025b).

Despite their diversity, most MT evaluation metrics, including Quality Estimation (QE) metrics that do not require references, share a structural property: they evaluate one candidate translation at a time and output an absolute scalar score (Rei et al., 2020, 2022; Guerreiro et al., 2024; Juraska et al., 2024). Typically, a metric conditions on the source segment, an optional reference translation, and a single candidate translation, and returns a scalar score. We posit that this single-candidate perspective is mismatched to several aspects of modern MT evaluation, especially in the high-quality settings where differences are subtle. In addition, one of the dominant downstream use cases is comparative, including model selection and candidate translations ranking (Kocmi et al., 2021, 2024c,a; Perrella et al., 2024). Motivated by evidence that comparative human judgments can be more consistent than absolute ratings in MT and related settings (Karpinska et al., 2021; Song et al., 2025), we ask whether supervised QE should treat pairwise comparison as the primary prediction task.

We present **PEAR** (Pairwise Evaluation for Au-

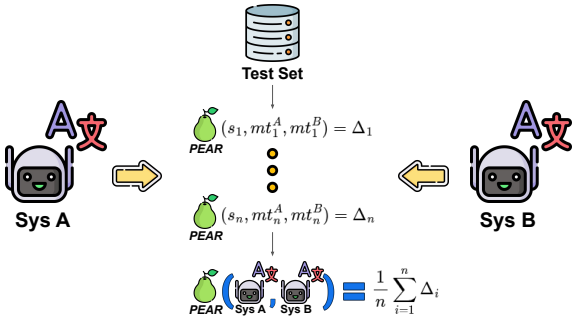

Figure 1: Segment- and system-level PEAR evaluation. For each source segment s_i, PEAR compares the system outputs (mt_i^A, mt_i^B) and predicts a relative score Δ_i. The sign indicates which translation is preferred ($\Delta_i > 0$: mt_i^A; $\Delta_i < 0$: mt_i^B), and values equal to zero correspond to ties; the magnitude $|\Delta_i|$ reflects the strength of the preference as a translation of s_i. The system-level PEAR score is then the arithmetic average of the segment-level PEAR scores.

tomatic Relative scoring), a supervised QE metric family that reframes reference-free MT evaluation as graded pairwise comparison. Given a source segment and two candidate translations, PEAR predicts the direction and magnitude of their quality difference. PEAR learns from pairwise supervision constructed by differencing human segment-level judgments under an antisymmetry-consistent objective, finally producing a relative score in output.

In summary, our contributions are:

- We introduce PEAR, a novel supervised QE formulation for MT evaluation based on graded pairwise relative scoring with antisymmetry-consistent training.
- We provide controlled comparisons showing that the proposed pairwise formulation improves over matched single-candidate QE under the same training setup, and we show that PEAR surpasses far larger QE submissions at WMT, while also outperforming strong reference-based metrics.
- We further show that PEAR can be applied without requiring N^2 system-to-system comparisons, and without relying on human reference translations, via a reference-anchored inference mode that remains effective when the anchor is an MT output.
- We analyze how PEAR relates to other metrics via inter-metric correlations, also demonstrating its effectiveness and efficiency as an MBR utility.

Figure 1 illustrates the PEAR execution flow at both the segment and system levels. We release PEAR, including trained checkpoints and a `pip`-installable package for easy use.[1]

2 Background and Related Work

Automatic MT evaluation metrics. Early MT metrics estimate quality via surface overlap against a reference translation, with BLEU (Papineni et al., 2002) and chrF (Popović, 2015) as widely used examples. Learned metrics based on pretrained representations improve agreement with human judgments by fine-tuning on human annotation signals, including reference-based models such as COMET and BLEURT (Rei et al., 2020; Sellam et al., 2020a,b) and more recent large-scale approaches such as XCOMET and MetricX (Guerreiro et al., 2024; Juraska et al., 2024). In parallel, QE metrics remove the dependency on references and score translations conditioned on the source and candidate only, including CometKiwi and its larger variants, as well as QE models of more recent learned metrics (Rei et al., 2022, 2023; Guerreiro et al., 2024; Juraska et al., 2024). LLM-based judging approaches, including GEMBA prompting for MQM or ESA, can be strong but are often expensive and may raise reproducibility concerns (Kocmi and Federmann, 2023b,a).

Pairwise and preference-based formulations. Pairwise evaluation has a long history in MT, including ranking-based training with engineered features and structured models (Ye et al., 2007; Duh, 2008; Guzmán et al., 2014, 2015). Several learned metrics leverage comparative supervision during training while still producing single-candidate scores at inference time, such as COMET-RANK (Rei et al., 2020) and reward-modeling approaches based on pairwise preferences (Tan and Monz, 2025). In contrast, COMET-poly incorporates additional context beyond the single translation at inference time, grounding the evaluation of a candidate in other candidates for the same source (Züfle et al., 2025). Closer to our inference setup, MT-RANKER formulates reference-free MT evaluation as binary classification: given a source and two candidates, it predicts which translation is better, so its output space does not represent ties or the strength of the preference (Moosa et al., 2024). PEAR differs by predicting graded relative differences without ruling out ties by design, and by directly targeting antisymmetric relative scoring.

Comparative human judgments. Recent human evaluation protocols increasingly use comparative setups, including side-by-side MQM for MT, motivated in part by improved consistency and agreement when annotators compare two candidates directly (Song et al., 2025). These findings align with broader observations that comparative judgments can reduce subjectivity in open-ended generation evaluation (Karpinska et al., 2021). PEAR draws on this motivation but addresses the automatic setting, treating comparative scoring as the prediction target for supervised QE in MT.

[1] Code and trained checkpoints will be made publicly available upon publication.

3 PEAR: Pairwise Evaluation for Automatic Relative scoring

In this section, we introduce PEAR, our pairwise framework for supervised reference-less MT evaluation. We begin by briefly recalling the standard regression-based formulation, where a metric assigns an absolute quality score to a single candidate translation. We then describe PEAR, which instead predicts a graded quality difference between two candidate translations of the same source text, together with the model architecture, training objective, and inference procedure used throughout the paper.

3.1 From Absolute to Relative Scoring

Most supervised MT metrics are trained to predict absolute scores: given a source segment s and a single candidate translation mt, the metric outputs an absolute scalar score. In reference-based evaluation, the metric is additionally provided with a reference translation r, whereas in QE settings it operates without r. Accordingly, these metrics take one of the following forms:

$$\hat{s}(s, mt) = g_\theta(s, mt),$$
$$\hat{s}(s, mt, r) = g_\theta(s, mt, r).$$

They are typically optimized to fit human segment-level judgments $s_h(s, mt)$[2] via a regression loss (Sellam et al., 2020a; Rei et al., 2020, 2022; Guerreiro et al., 2024; Juraska et al., 2024);[3] other work casts automatic MT evaluation as reward modeling, learning from human preferences while still producing absolute scores for individual candidate translations at inference time (Tan and Monz, 2025). When comparing two candidates mt_a and mt_b, an implicit relative score can then be obtained by subtraction, i.e., by computing $\hat{s}(s, mt_a) - \hat{s}(s, mt_b)$ in reference-less evaluation, or $\hat{s}(s, mt_a, r) - \hat{s}(s, mt_b, r)$ in reference-based evaluation.

Rather than deriving a comparison by subtracting two independently predicted absolute scores, an alternative approach is to predict the preference directly from a joint encoding of the source and the two candidate translations. MT-RANKER (Moosa et al., 2024) follows this approach in the QE setting, taking (s, mt_a, mt_b) as input and predicting which candidate translation is better through a binary classification framing. However, this formulation collapses comparison to a two-way decision: it cannot quantify the strength of the preference, and it rules out ties by design, since ties are not part of the output label space. PEAR instead treats comparison as the prediction target: given a source text, it scores a pair of candidate translations in a shared input context, producing a graded preference directly instead of relying on differences of separately predicted absolute scores.

3.2 Problem Formulation

Let s be a source segment and let mt_a, mt_b be two candidate translations of s. A PEAR model f_θ predicts a real-valued relative score

$$\hat{\Delta}_{ab} = f_\theta(s, mt_a, mt_b) \in \mathbb{R},$$

interpreted as the predicted quality difference between mt_a and mt_b as translations of s. Positive values favor mt_a, negative values favor mt_b, and a value of zero indicates a tie.

PEAR is trained to approximate human pairwise differences. Specifically, given human segment-level scores $s_h(s, mt)$, we construct supervision for a candidate pair as

$$\Delta^\star_{ab} = s_h(s, mt_a) - s_h(s, mt_b).$$

At the system level, let $\{s_i\}_{i=1}^n$ be a test set and let $\{(mt_i^A, mt_i^B)\}_{i=1}^n$ be the corresponding output pairs produced by two MT systems S_A and S_B. PEAR produces segment-level predictions $\hat{\Delta}_i = f_\theta(s_i, mt_i^A, mt_i^B)$, which are aggregated by arithmetic mean to yield a system-level graded preference:

$$\hat{\Delta}(S_A, S_B) = \frac{1}{n}\sum_{i=1}^{n} \hat{\Delta}_i,$$

as also illustrated in Figure 1.

3.3 Model Architecture and Input Formatting

PEAR adopts a cross-encoder setup that jointly encodes the source text and the two candidate translations. We instantiate the encoder with InfoXLM Large[4] for PEAR (Chi et al., 2021) and with XLM-RoBERTa-XL[5] for PEAR-XL (Goyal et al., 2021).

[2] Several recent human evaluation protocols for MT are reference-free: annotators assess system outputs only with respect to the source text, which reduces dependence on any particular reference and thereby mitigates reference bias (Freitag et al., 2021; Kocmi et al., 2022, 2024b).

[3] In practice, this regression loss is commonly implemented as Mean Squared Error (MSE) between the predicted scores and the human judgments.

[4] https://huggingface.co/microsoft/infoxlm-large

[5] https://huggingface.co/facebook/xlm-roberta-xl

Throughout, we describe the scoring computation for a single input sample, from input construction to the final relative score.

Input Serialization. Given a source segment s and two candidate translations mt_a and mt_b, we serialize them into a single sequence:

$$x \;=\; [\texttt{BOS}]\; s\; [\texttt{SEP}]\; mt_a\; [\texttt{SEP}]\; mt_b\; [\texttt{EOS}],$$

where [BOS], [SEP], and [EOS] denote the encoder's special tokens.[6] We additionally build binary span masks $\mathbf{m}_{\text{src}}, \mathbf{m}_a, \mathbf{m}_b \in \{0,1\}^T$ over token positions in x that select the content tokens of s, mt_a, and mt_b.[7]

Encoder Representations. We denote by

$$\mathbf{H} = \text{Enc}_\phi(x) \in \mathbb{R}^{T\times d}$$

the contextual token representations extracted from the encoder's final layer. We apply masked mean pooling to obtain span representations:

$$\mathbf{h}_\ell \;=\; \frac{\sum_{t=1}^{T} m_{\ell,t}\,\mathbf{H}_t}{\max\!\left(\sum_{t=1}^{T} m_{\ell,t},\, \varepsilon\right)} \quad \text{for } \ell \in \{\text{src}, a, b\},$$

where $\mathbf{H}_t \in \mathbb{R}^d$ is the token vector at position t and $\varepsilon > 0$ avoids division by zero.

Pairwise Head. We construct a source-aware representation for each candidate translation $k \in \{a, b\}$:

$$\boldsymbol{\varphi}_k = \left[\mathbf{h}_k;\; \mathbf{h}_k \odot \mathbf{h}_{\text{src}};\; |\mathbf{h}_k - \mathbf{h}_{\text{src}}|\right] \in \mathbb{R}^{3d},$$

where $\odot$ and $|\cdot|$ are applied elementwise. Using shared parameters, we map $\boldsymbol{\varphi}_k$ to a scalar utility term:

$$\mathbf{a}_k = \text{Proj}(\boldsymbol{\varphi}_k) \in \mathbb{R}^d,$$
$$u_k = \text{FFN}(\mathbf{a}_k) \in \mathbb{R}.$$

Where both Proj and FFN are linear layers with dropout and GELU activation (Hendrycks and Gimpel, 2016) in the middle. We then form a comparison logit by subtraction:

$$z \;=\; u_a - u_b.$$

Note that only the difference is used for supervision and inference, so the individual u_k values are not intended as absolute quality scores. The final output is a scaled relative score:

$$\hat{\Delta}_{ab} \;=\; \alpha\, z, \qquad \alpha \;=\; \text{softplus}(\alpha_{\text{raw}}) + \varepsilon,$$

where $\alpha_{\text{raw}} \in \mathbb{R}$ is a learned scalar parameter. The softplus reparameterization, together with the addition of a small constant $\varepsilon > 0$, ensures $\alpha > 0$.

3.4 Training Objective

A key property of relative scoring is antisymmetry: swapping the candidates should negate the predicted difference while preserving its magnitude, i.e., $f_\theta(s, mt_a, mt_b) = -f_\theta(s, mt_b, mt_a)$. To encourage this behavior, for each training instance (s, mt_a, mt_b) we also consider the swapped order (s, mt_b, mt_a) and denote the corresponding prediction by $\hat{\Delta}_{ba} = f_\theta(s, mt_b, mt_a)$. Specifically, we train with a Huber loss (Huber, 1964) on the human difference together with a flip-consistency regularizer:

$$\mathcal{L}_{\text{diff}} = \ell_\delta\!\left(\hat{\Delta}_{ab} - \Delta^\star_{ab}\right),$$
$$\mathcal{L}_{\text{flip}} = \left(\hat{\Delta}_{ab} + \hat{\Delta}_{ba}\right)^2,$$
$$\mathcal{L} = \mathcal{L}_{\text{diff}} + \lambda_{\text{flip}}\,\mathcal{L}_{\text{flip}},$$

where λ_{flip} is a hyperparameter controlling the strength of the antisymmetry constraint. Here ℓ_δ denotes the Huber loss (Huber, 1964):

$$\ell_\delta(r) \;=\; \begin{cases} \frac{1}{2}r^2 & \text{if } |r| \le \delta, \\ \delta\left(|r| - \frac{1}{2}\delta\right) & \text{otherwise,} \end{cases}$$

where $\delta > 0$ is a hyperparameter that sets the residual magnitude at which the penalty transitions from quadratic to linear.[8]

3.5 Inference Modes

We use PEAR in two inference configurations. In its default pairwise QE mode, it compares two candidate translations of the same source segment without access to reference translations. When a human reference is available, we additionally consider a reference-anchored mode by fixing one side of the pair to the reference.

Pairwise QE Mode (PEAR). Given a source segment s and two candidate translations mt_a and mt_b, PEAR returns a relative score:

$$\hat{\Delta}_{ab} \;=\; f_\theta(s, mt_a, mt_b).$$

[6] For InfoXLM and XLM-R encoders, these correspond to <s>, </s></s>, and </s>, respectively.

[7] Special tokens are masked out in this process.

[8] Compared to MSE, the Huber loss is less sensitive to large residuals (Huber, 1964); in Appendix A, we report an ablation where it improves performance over MSE in our setting.

To reduce sensitivity to the candidates' order, we optionally also score the swapped order and combine the two predictions:

$$\hat{\Delta}_{ba} = f_\theta(s, mt_b, mt_a),$$
$$\tilde{\Delta}_{ab} = \tfrac{1}{2}\big(\hat{\Delta}_{ab} - \hat{\Delta}_{ba}\big).$$

Reference-Anchored Mode (PEAR$_{\text{ref}}$). When a human reference translation r is available, we anchor the comparison by fixing one side of the input to the reference:

$$\hat{\Delta}(mt, r) = f_\theta(s, mt, r).$$

This does not make PEAR a reference-based metric in the usual sense; rather, it instantiates the same relative-scoring function in a reference-anchored configuration. As above, we can also score the swapped order $f_\theta(s, r, mt)$ and combine the two predictions with the same order-combination rule.

A practical advantage of this reference-anchored configuration is computational. When evaluating N systems on an evaluation set $\{(s_i, r_i)\}_{i=1}^n$ with system outputs $\{mt_i^{(j)}\}_{i=1}^n$ for each system S_j, it yields one reference-anchored score per system:

$$\hat{\Delta}(S_j, r) = \frac{1}{n}\sum_{i=1}^{n} f_\theta(s_i, mt_i^{(j)}, r_i).$$

This requires $O(N)$ system scores, avoiding the $N(N-1)/2$ system-to-system comparisons (quadratic in N) needed by fully pairwise evaluation.

4 Experimental Setup

We now summarize the empirical setting for our main experiments. We first describe the training and evaluation procedures (Section 4.1) and report the trained PEAR models (Section 4.2). Baselines, data statistics, and training hyperparameters are reported in the Appendix B.

4.1 Training and Evaluation Data

WMT Supervision. We train PEAR on WMT human evaluation data, converting absolute segment-level human assessments into pairwise supervision via score differencing (Section 3.2). Our training data combines DA, DA+SQM, and MQM annotations, with MQM offering the most fine-grained supervision. Following common practices in MT metrics training, we adopt a two-stage schedule (Guerreiro et al., 2024; Juraska et al., 2024). In the first training stage, we pre-train on DA and DA+SQM judgments released in the WMT evaluation campaigns from WMT16 to WMT23, providing broad coverage across language directions and translation quality. In the second training stage, we fine-tune exclusively on MQM supervision from WMT20 to WMT23,[9] additionally including the IndicMT Eval MQM dataset for English→Indic directions (Sai B et al., 2023), which has been used to train XCOMET (Guerreiro et al., 2024), a metric included in our comparisons.

Scaling via Distilled MQM Supervision. MQM gold data are available for relatively few language pairs and are expensive to collect. To stress-test whether the proposed pairwise QE framing scales favorably with additional supervision, we also test PEAR models whose second training stage is augmented with MQM annotations distilled from GPT-4.1-mini on language pairs not covered by MQM, prompting it with a GEMBA-MQM V2 approach (Junczys-Dowmunt, 2025). Additional details on these data are reported in Appendix B.

Evaluation Benchmark. Our primary benchmark is the MQM test set released with the WMT24 Metrics Shared Task. We adopt the official WMT Metrics Shared Task evaluation toolkit.[10] Following the WMT24 setup, we report Soft Pairwise Accuracy (SPA) at the system level (Thompson et al., 2024) and pairwise accuracy with tie calibration (acc^*_{eq}) at the segment level (Deutsch et al., 2023). Since PEAR outputs pairwise scores by design, we interface it with the toolkit by producing segment-level scores for each system pair.

4.2 Trained Models

PEAR Models. We train and evaluate two PEAR variants: **PEAR**, instantiated with InfoXLM Large (Chi et al., 2021), and **PEAR-XL**, instantiated with XLM-RoBERTa-XL (Goyal et al., 2021).

Matched Absolute-Scoring Baselines. To disentangle the effect of the proposed pairwise QE formulation from backbone capacity and training data exposure, we also train matched absolute-scoring QE baselines that take only the source segment and a single candidate translation as input, share

[9] This is done not only because MQM offers the most fine-grained MT quality assessment, but also to align the training signal with our main target evaluation setting, since WMT24 meta-evaluation is centered on MQM.

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

Group	Metric	θ	Ref?	SPA	acc^*_{eq}	Avg Corr
Single Candidate	Single-QE-XL$_{\text{KD}}$	3.5B	×	80.9	57.9	69.4
	Single-QE-XL	3.5B	×	80.4	57.6	69.0
	Single-QE$_{\text{KD}}$	560M	×	80.6	57.4	69.0
	Single-QE	560M	×	80.0	57.2	68.6
Pairwise (PEAR)	PEAR-XL$_{\text{both,KD}}$	3.5B	×	**82.1**	58.1	**70.1**
	PEAR-XL$_{\text{KD}}$	3.5B	×	82.0	**58.2**	**70.1**
	PEAR-XL$_{\text{both}}$	3.5B	×	81.4	**58.2**	69.8
	PEAR-XL	3.5B	×	81.5	58.1	69.8
	PEAR$_{\text{both,KD}}$	560M	×	81.9	58.1	70.0
	PEAR$_{\text{KD}}$	560M	×	81.8	**58.2**	70.0
	PEAR$_{\text{both}}$	560M	×	81.2	58.0	69.6
	PEAR	560M	×	80.9	57.9	69.4

Table 1: Controlled comparison of PEAR against matched single-candidate QE baselines on the WMT24 MQM test set. SPA and acc^*_{eq} are averaged over En-De, En-Es, and Ja-Zh; Avg Corr is the mean of these two averages. Bold indicates the best score in each column. For PEAR, *both* denotes the bidirectional pairwise QE configuration, computed by averaging the two relative scores from both input orders. KD denotes models fine-tuned with additional MQM supervision distilled from GPT-4.1-mini.

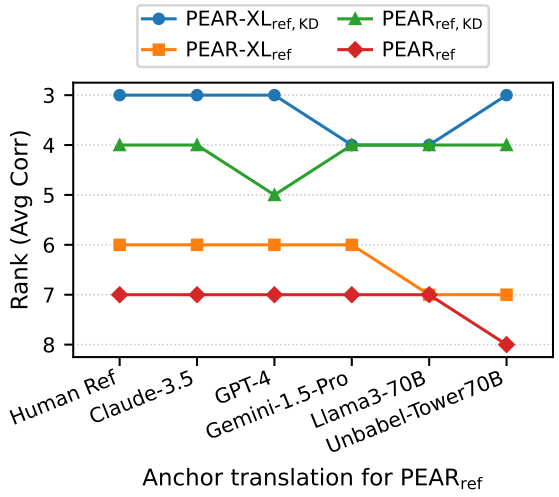

Figure 2: Rank stability of PEAR$_{\text{ref}}$ across anchors. The leftmost anchor (*Human Ref*) uses the human reference as the fixed comparison translation, matching Table 4. The remaining anchors use an MT output; for each MT anchor, that system is removed from the benchmark before recomputing meta-evaluation. Ranks are computed by Avg Corr (lower is better).

Utility	LP	XCOMET-XL	CometKiwi-XL	MetricX-XL
PEAR (full)	En-De	**0.855**	**0.731**	**-5.2**
PEAR (sym.)	En-De	0.854	**0.731**	**-5.4**
COMET	En-De	0.844	0.730	-6.6
BLEURT-20	En-De	0.842	0.728	-6.3
PEAR (full)	En-Ja	**0.810**	**0.685**	**-6.4**
PEAR (sym.)	En-Ja	0.809	**0.685**	**-6.7**
COMET	En-Ja	0.798	0.684	-7.8
BLEURT-20	En-Ja	0.796	0.683	-6.9

Table 2: MBR decoding on 100-best lists for WMT24 En→De and En→Ja. PEAR (sym.) computes only one triangle of the $N \times N$ utility matrix and fills the remainder by antisymmetry.

the same encoder backbones and training data, and predict an absolute scalar quality score for that candidate. These QE baselines let us validate the proposed pairwise QE formulation at the methodological level, testing it as an alternative to conventional absolute-score QE under matched training conditions.

5 Results and Analyses

This section reports empirical evidence for PEAR across several settings. We start with reporting results on WMT24 (Section 5.1 and Section 5.2), and continue by showing additional analyses on the behavior of PEAR models (Section 5.3, Section 5.4, and Section 5.5).

5.1 Pairwise QE vs. Single-Candidate QE at Scale

This subsection isolates the effect of the proposed pairwise QE framing from confounding factors such as backbone capacity, training data exposure, and model selection. To that end, we compare PEAR against strictly matched single-candidate QE baselines trained with the same data, hyperparameters, and backbone model. Checkpoint selection is performed on WMT23 MQM data (held out from training), and results are reported on the WMT24 MQM benchmark.

We report both PEAR in its default configuration (one input order) and the bidirectional variant that combines predictions from both input orders, matching the inference mode described in Section 3.5. We further consider settings with and

without knowledge distillation augmentation in the second training stage, to probe whether the advantage of our pairwise QE formulation persists—and potentially widens—as the amount of MQM supervision is scaled up via distilled annotations.

Table 1 reports the comparison results. Across backbones and training regimes, PEAR yields better performance than the matched single-candidate baselines, indicating that, with the same training data and encoder backbone, PEAR's pairwise QE framing improves over single-candidate QE for comparing candidate translations.

5.2 WMT24 Results

Table 4 reports results on the WMT24 meta-evaluation benchmark. For PEAR, we report only the bidirectional configuration (*both*), since Table 1 shows that single-pass and *both* yield very similar performance. All PEAR variants in Table 4 are trained with the same hyperparameters used in the controlled comparison of Table 1, with the only difference that we also include WMT23 data in training.

Better Performance with Fewer Parameters. Among reference-free metrics, PEAR attains high Avg Corr with substantially smaller models. PEAR-XL$_{\text{both}}$ (3.5B) exceeds MetricX-24-Hybrid-QE-XL (3.7B) and XCOMET-QE (24B) on Avg Corr (70.2 vs. 69.9 and 69.5), while using roughly 7$\times$ fewer parameters than XCOMET-QE. Distilled supervision yields a further improvement, raising Avg Corr to 70.5.

The same trend holds with the smaller PEAR models. PEAR$_{\text{both}}$ (560M) remains higher than MetricX-24-Hybrid-QE-XL (3.7B) and XCOMET-QE (24B) on Avg Corr (70.1 vs. 69.9 and 69.5), despite using about 7$\times$ and 40$\times$ fewer parameters, respectively. The only QE metric in Table 4 that is comparable in size to PEAR is CometKiwi (560M), yet it achieves a markedly lower Avg Corr (64.0 vs. 70.1). With distilled supervision, PEAR$_{\text{both,KD}}$ also edges out CometKiwi-XXL on Avg Corr (70.4 vs. 70.3), while using about 20$\times$ fewer parameters (560M vs. 10.5B).

Comparison against strong reference-based metrics. Even without references, PEAR compares favorably to strong reference-based metrics. For example, PEAR-XL$_{\text{both,KD}}$ matches the Avg Corr of MetricX-24-Hybrid-Large (70.5 vs. 70.5), and PEAR$_{\text{both}}$ exceeds COMET-22 and BLEURT-20 (70.1 vs. 68.9 and 68.6).

5.3 Does PEAR$_{\text{ref}}$ Require Human References?

Table 4 shows that the reference-anchored configuration, PEAR$_{\text{ref}}$ (Section 3.5), slightly outperforms its corresponding QE (*both*) variant on WMT24. We now test whether this advantage depends on anchoring to a human reference, or whether the same behavior holds when the anchor is an MT output.

We instantiate the anchor slot with (i) the human reference (*Human Ref*) and (ii) the output of five MT systems. For each MT anchor, we re-run WMT24 meta-evaluation after removing the anchor system from the benchmark, since its outputs are used as the fixed comparison target.[11]

Figure 2 shows that the ranks of PEAR$_{\text{ref}}$ variants are highly stable across anchors. Replacing the human reference with an MT output almost never changes the rank of PEAR$_{\text{ref}}$, and when changes occur, they shift by at most one rank position. This suggests that PEAR$_{\text{ref}}$ does not rely on human references to retain its performance: anchoring to MT outputs provides a similar reference point, making the reference-anchored interface practical even when human references are unavailable.

5.4 Correlation Between PEAR and Other Metrics

PEAR is trained with a pairwise relative-scoring objective (Section 3), in contrast to the single-candidate regression objective used by most supervised WMT24 metrics. To assess how distinct PEAR's evaluation signal is, we analyze its segment-level correlation with other metrics.

For this analysis, we compute the Pearson correlation between pairwise segment-level difference scores. Since PEAR outputs pairwise scores directly, we use its predicted differences as-is. For metrics that output single-candidate segment scores, we convert them into pairwise difference scores by subtraction for each segment and MT system pair, i.e., $\Delta_m(s, mt_a, mt_b) = m(s, mt_a) - m(s, mt_b)$. We compute correlations separately for each WMT24 MQM language pair.

Figures 3, 4, and 5 show the resulting correlation matrices for En-De, En-Es, and Ja-Zh. Across all three language pairs, PEAR$_{\text{both}}$ and PEAR$_{\text{both,KD}}$ have consistently lower correlation with the other

[11] Since the set of evaluated systems changes with the chosen MT anchor, the resulting Avg Corr values are not directly comparable across anchors. We therefore report ranks (by Avg Corr) rather than raw Avg Corr values.

strong WMT24 metrics. For example, their correlation with MetricX-24-Hybrid-QE[12] is moderate on En-De ($r \approx 0.71$), drops on En-Es ($r \approx 0.51$), and is much lower on Ja-Zh ($r \approx 0.26$). Excluding lexical and unsupervised baselines (BLEU, chrF, and BERTScore), PEAR metrics are the least correlated with the rest of the metric suite across all three language pairs. Overall, this may suggest that PEAR captures different evaluation signals relative to existing WMT24 metrics. We leave a deeper investigation of the sources and implications of these lower correlations to future work.

5.5 PEAR for MBR Decoding

Using reference-based metrics such as COMET and BLEURT-20 for MBR decoding repurposes models trained to condition on human references, which can be a mismatch when utilities are computed purely between candidate translations. PEAR, in contrast, is trained explicitly for reference-free pairwise comparison and does not assume that either candidate serves as a reference. In this section, we compare PEAR, COMET, and BLEURT-20 as utility functions for MBR decoding.

Candidate lists. We translate the English source side of the WMT24 test set into German and Japanese with a transformer-based multilingual MT system trained on public and internal data with teacher-student knowledge distillation. We decode with beam search and retain a 100-best list for each source segment.

MBR utilities. We run MBR over each 100-best list using three utilities: PEAR,[13] COMET, and BLEURT-20. For PEAR, we compare two implementations. The first computes the full utility matrix, while the second exploits antisymmetry and evaluates only one triangle of the matrix, setting $u(h_j, h_i) = -u(h_i, h_j)$, halving the number of forward passes. This tests whether PEAR behaves as an approximately antisymmetric utility during decoding, while enabling a cheaper MBR procedure.

Results. We evaluate the resulting MBR outputs with three strong metrics: XCOMET-XL, CometKiwi-XL, and MetricX-24-Hybrid-XL. Table 2 shows that PEAR yields nearly identical performance under the full and symmetry-reduced MBR implementations, indicating that the symmetry shortcut is effective in practice. Moreover, PEAR-based MBR improves over COMET and BLEURT-20 under XCOMET-XL and MetricX-24-Hybrid-XL, while gains under CometKiwi-XL are relatively smaller. These results suggest that PEAR, trained explicitly for pairwise comparison, can be a stronger utility for MBR decoding than traditional reference-based metrics.

[12] It corresponds to the XXL checkpoint.

[13] InfoXLM Large checkpoint fine-tuned without distilled supervision. We use the default inference configuration, not the *both* variant.

6 Conclusion

We presented PEAR, a supervised QE metric family that models MT evaluation as graded pairwise relative scoring. Unlike standard metrics that score one candidate at a time, PEAR compares two translations jointly and predicts both the direction and strength of preference. In controlled experiments, PEAR consistently outperforms matched single-candidate QE baselines, indicating that the pairwise formulation is a better method for comparing candidate translations.

On WMT24 MQM meta-evaluation, PEAR attains higher correlation than the largest QE submissions, also outperforming strong reference-based metrics despite operating in a reference-free setting. We also found that PEAR_{ref} remains effective when the anchor translation is produced by a strong MT system rather than a human reference, and that PEAR's segment-level pairwise scores are less correlated with other top metrics, suggesting that PEAR captures a different evaluation signal. Future work will focus on understanding which phenomena drive this divergence, and on leveraging pairwise relative scoring in other evaluation settings.

7 Limitations

Model scale. Our experiments do not fully characterize how far PEAR can be pushed with substantially larger models. The largest PEAR checkpoint we fine-tune in this work is PEAR-XL (3.5B parameters), and we therefore do not test whether the performance gains we attribute to the pairwise formulation persist, widen, or saturate with larger models. Recent works suggest that scaling up the underlying model and fine-tuning larger backbones can be a strong driver of performance for supervised metrics (Rei et al., 2023; Guerreiro et al., 2024; Juraska et al., 2024, 2025; Tan and Monz, 2025). Extending PEAR to larger model sizes is a natural next step.

Metric divergence. In Section 5.4, PEAR exhibits lower correlation with other strong WMT24 metrics than is typical among top-performing metrics. In this work, we treat this primarily as evidence that PEAR may induce a complementary evaluation signal, but we do not deeply investigate how this could be exploited. For example, lower correlation could be beneficial when combining metrics in an ensemble, or when other metrics are used as optimization targets, such as rewards in reinforcement learning, since it may reduce over-optimization to a single signal and support fairer automatic assessment, but it could also indicate sensitivity to specific phenomena that warrant targeted diagnosis. A more detailed analysis of this divergence is left for future work.

Targeted synthetic data. Unlike some recent metrics that incorporate targeted synthetic examples designed to address known failure modes (e.g., fluent but unrelated translations, undertranslation, and related specific errors), PEAR is not fine-tuned with comparable hand-designed perturbation suites (Juraska et al., 2023, 2024; Guerreiro et al., 2024). Although we experiment with scaling supervision via distilled MQM annotations, we do not study whether PEAR particularly benefits from synthetic pair construction that is explicitly contrastive and pairwise by design. PEAR's pairwise interface makes targeted contrastive pair construction more natural and potentially more cost-effective than for other traditional supervised metrics. We leave a systematic investigation of such targeted synthetic data strategies, and their interaction with pairwise relative-score training, for future work.

References

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An Information-Theoretic Framework for Cross-Lingual Language Model Pre-Training. in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

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, Singapore. Association for Computational Linguistics.

Kevin Duh. 2008. Ranking vs. Regression in Machine Translation Evaluation. in *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 191–194, Columbus, Ohio. Association for Computational Linguistics.

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.

Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. 2021. Larger-Scale Transformers for Multilingual Masked Language Modeling. in *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 29–33, Online. Association for Computational Linguistics.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xCOMET: Transparent Machine Translation Evaluation through Fine-grained Error Detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, Alessandro Moschitti, Preslav Nakov, and Massimo Nicosia. 2014. Learning to Differentiate Better from Worse Translations. in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 214–220, Doha, Qatar. Association for Computational Linguistics.

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2015. Pairwise Neural Machine Translation Evaluation. in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 805–814, Beijing, China. Association for Computational Linguistics.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian Error Linear Units (GELUs). *arXiv preprint arXiv:1606.08415*.

Peter J. Huber. 1964. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73–101.

Marcin Junczys-Dowmunt. 2025. GEMBA V2: Ten Judgments Are Better Than One. in *Proceedings of*

the Tenth Conference on Machine Translation, pages 926–933, Suzhou, China. Association for Computational Linguistics.

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, Miami, Florida, USA. Association for Computational Linguistics.

Juraj Juraska, Tobias Domhan, Mara Finkelstein, Tetsuji Nakagawa, Geza Kovacs, Daniel Deutsch, Pidong Wang, and Markus Freitag. 2025. MetricX-25 and GemSpanEval: Google Translate Submissions to the WMT25 Evaluation Shared Task. in *Proceedings of the Tenth Conference on Machine Translation*, pages 957–968, Suzhou, China. Association for Computational Linguistics.

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, Singapore. Association for Computational Linguistics.

Marzena Karpinska, Nader Akoury, and Mohit Iyyer. 2021. The Perils of Using Mechanical Turk to Evaluate Open-Ended Text Generation. in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1265–1285, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024a. 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, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 Conference on Machine Translation (WMT22). in *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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, Singapore. Association for Computational Linguistics.

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, Tampere, Finland. European Association for Machine Translation.

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, Online. Association for Computational Linguistics.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024b. Error Span Annotation: A Balanced Approach for Human Evaluation of Machine Translation. in *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Vilém Zouhar, Christian Federmann, and Matt Post. 2024c. Navigating the Metrics Maze: Reconciling Score Magnitudes and Accuracies. in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1999–2014, Bangkok, Thailand. Association for Computational Linguistics.

Benjamin Marie, Atsushi Fujita, and Raphael Rubino. 2021. Scientific Credibility of Machine Translation Research: A Meta-Evaluation of 769 Papers. in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7297–7306, Online. Association for Computational Linguistics.

Ibraheem Muhammad Moosa, Rui Zhang, and Wenpeng Yin. 2024. MT-Ranker: Reference-free machine translation evaluation by inter-system ranking. in *The Twelfth International Conference on Learning Representations*.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Stefano Perrella, Lorenzo Proietti, Pere-Lluís Huguet Cabot, Edoardo Barba, and Roberto Navigli. 2024. Beyond Correlation: Interpretable Evaluation of Machine Translation Metrics. in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20689–20714,

Miami, Florida, USA. Association for Computational Linguistics.

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, Lisbon, Portugal. Association for Computational Linguistics.

Lorenzo Proietti, Stefano Perrella, and Roberto Navigli. 2025a. Has Machine Translation Evaluation Achieved Human Parity? The Human Reference and the Limits of Progress. in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 790–813, Vienna, Austria. Association for Computational Linguistics.

Lorenzo Proietti, Stefano Perrella, Vilém Zouhar, Roberto Navigli, and Tom Kocmi. 2025b. Estimating Machine Translation Difficulty. in *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 24261–24285, Suzhou, China. Association for Computational Linguistics.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André F. T. Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 Submission for the Quality Estimation Shared Task. in *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. Association for Computational Linguistics.

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)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-Unbabel 2022 Submission for the Quality Estimation Shared Task. in *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ananya Sai B, Tanay Dixit, Vignesh Nagarajan, Anoop Kunchukuttan, Pratyush Kumar, Mitesh M. Khapra, and Raj Dabre. 2023. IndicMT Eval: A Dataset to Meta-Evaluate Machine Translation Metrics for Indian Languages. in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14210–14228, Toronto, Canada. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020a. BLEURT: Learning Robust Metrics for Text Generation. in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Thibault Sellam, Amy Pu, Hyung Won Chung, Sebastian Gehrmann, Qijun Tan, Markus Freitag, Dipanjan Das, and Ankur Parikh. 2020b. Learning to Evaluate Translation Beyond English: BLEURT Submissions to the WMT Metrics 2020 Shared Task. in *Proceedings of the Fifth Conference on Machine Translation*, pages 921–927, Online. Association for Computational Linguistics.

Yixiao Song, Parker Riley, Daniel Deutsch, and Markus Freitag. 2025. Enhancing Human Evaluation in Machine Translation with Comparative Judgement. in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20536–20551, Vienna, Austria. Association for Computational Linguistics.

Shaomu Tan and Christof Monz. 2025. ReMedy: Learning Machine Translation Evaluation from Human Preferences with Reward Modeling. in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 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, Miami, Florida, USA. Association for Computational Linguistics.

Yang Ye, Ming Zhou, and Chin-Yew Lin. 2007. Sentence Level Machine Translation Evaluation as a Ranking. in *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 240–247, Prague, Czech Republic. Association for Computational Linguistics.

Vilém Zouhar, Tom Kocmi, and Mrinmaya Sachan. 2025a. AI-Assisted Human Evaluation of Machine Translation. in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4936–4950, Albuquerque, New Mexico. Association for Computational Linguistics.

Vilém Zouhar, Peng Cui, and Mrinmaya Sachan. 2025b. How to Select Datapoints for Efficient Human Evaluation of NLG Models? *Preprint*, arXiv:2501.18251.

Maike Züfle, Vilém Zouhar, Tu Anh Dinh, Felipe Maia Polo, Jan Niehues, and Mrinmaya Sachan. 2025. COMET-poly: Machine Translation Metric Grounded in Other Candidates. in *Proceedings of the Tenth Conference on Machine Translation*, pages 887–904, Suzhou, China. Association for Computational Linguistics.

Loss	SPA ↑	acc^*_{eq} ↑	Avg Corr ↑
MSE	80.6	57.6	69.1
Huber	**80.9**	**57.9**	**69.4**

Table 3: WMT24 meta-evaluation for PEAR (InfoXLM Large, 560M) trained with MSE vs. Huber regression on pairwise human-difference supervision. SPA is reported at the system level (Thompson et al., 2024); acc^*_{eq} is segment-level pairwise accuracy with tie calibration (Deutsch et al., 2023).

A Huber Loss vs. MSE for Pairwise Difference Regression

PEAR is trained to regress human quality differences $\Delta^\star_{ab}$ (Section 3.4). In all the experiments, we use the Huber loss, which is less sensitive to occasional large residuals than MSE (Huber, 1964).

To quantify the impact of this choice, we run an ablation for PEAR (InfoXLM Large, 560M) in the same training setup as Section 5.1, changing only the regression loss from Huber to MSE while keeping all other hyperparameters fixed. We evaluate on WMT24 MQM with the official meta-evaluation toolkit, reporting Soft Pairwise Accuracy (SPA) at the system level (Thompson et al., 2024), pairwise accuracy with tie calibration (acc^*_{eq}) at the segment level (Deutsch et al., 2023), and Avg Corr.

Table 3 shows that Huber yields small but consistent gains over MSE across meta-evaluation statistics. We hypothesize that down-weighting large residuals stabilizes learning under heavy-tailed pairwise targets, which can improve calibration for close comparisons among strong MT systems. Further analysis of when and why these gains arise is left to future work.

B Training and Evaluation Setup

We report the results discussed in Section 5.2 in Table 4.

B.1 Data

PEAR training set is composed of 7M translation pairs across 51 language pairs for the first training stage (DA and DA+SQM), and by 3M across 10 language pairs for the second training stage (MQM). We do not apply any normalization to the scores provided by human annotators, following Juraska et al. (2024). For the KD data, we run the GEMBA-MQM V2 approach on WMT data annotated only with DA and DA+SQM, and on internal MT output data, featuring a total of 1M additional translation pairs, covering 5 additional language pairs in addition to those present in gold MQM data.

B.2 Hyperparameters

For the experiments outlined in Section 5.1, we train all the models with AdamW with a learning rate equal to 2e-5 until convergence on the development set (WMT23 MQM data). Then, for the experiments described in Section 5.2 we use the same hyperparameters, but also include WMT23 in the training set, and train for only one epoch. The λ_{flip} hyperparameter has been set to 0.1 after preliminary hyperparameter optimization experiments, where we did not observe large differences in the range [0.1, 0.5], with a step size of 0.1. The δ hyperparameter has been set to 4.5 after preliminary hyperparameter optimization experiments, where we did not observe large differences in the range [2.0, 8.0], with a step of 0.1. The parameter α_{raw} presented in Section 3 is kept frozen for the first training stage on DA and DA+SQM data, and it is learned in the second training stage on MQM, with an initial value of 1.0. This is done in order to allow the model to more easily adapt to the different training data distribution in the second training stage, coming from MQM data, not DA or DA+SQM.

B.3 Baselines

Since PEAR targets improvements in QE, our primary comparisons are against top-performing QE metric families submitted as primary entries to the WMT24 Metrics Shared Task. For additional context, we also report a small set of strong reference-based baselines.

B.3.1 QE metrics.

COMET family (QE). We include CometKiwi (Rei et al., 2022), its large-scale variant CometKiwi-XXL (Rei et al., 2023), and XCOMET-QE (Guerreiro et al., 2024).

MetricX family. We include the QE variants of MetricX-24 Hybrid (Juraska et al., 2024), namely MetricX-24-Hybrid-QE-Large, -XL, and -XXL.

GEMBA family. We include GEMBA-ESA, a WMT24 primary submission that prompts GPT-4 to follow the ESA procedure by extracting error spans and then producing a final 0–100 score (Kocmi and Federmann, 2023a,b; Freitag et al., 2024).

Group	Metric	θ	Ref?	SPA	acc^*_{eq}	Avg Corr
WMT24 Metrics	MetricX-24-Hybrid-QE-XXL	13B	×	84.9	58.0	**71.4**
	MetricX-24-Hybrid-QE-XL	3.7B	×	83.4	56.5	69.9
	MetricX-24-Hybrid-QE-Large	1.2B	×	80.6	56.1	68.3
	XCOMET-QE	24B‡	×	83.3	55.7	69.5
	CometKiwi-XXL	10.5B	×	**85.4**	55.2	70.3
	CometKiwi	560M	×	73.3	54.7	64.0
	GEMBA-ESA	GPT-4	×	84.6	57.6	71.1
	MetricX-24-Hybrid-Large	1.2B	✓	84.0	57.0	70.5
	COMET-22	560M	✓	82.4	55.4	68.9
	BLEURT-20	579M	✓	82.1	55.0	68.6
Ours	PEAR-XL$_{\text{ref,KD}}$	3.5B	✓†	82.7	58.9	70.8
	PEAR-XL$_{\text{both,KD}}$	3.5B	×	82.7	58.3	70.5
	PEAR-XL$_{\text{ref}}$	3.5B	✓†	81.6	**59.2**	70.4
	PEAR-XL$_{\text{both}}$	3.5B	×	81.6	58.8	70.2
	PEAR$_{\text{ref,KD}}$	560M	✓†	82.6	58.7	70.7
	PEAR$_{\text{both,KD}}$	560M	×	82.0	58.8	70.4
	PEAR$_{\text{ref}}$	560M	✓†	81.5	59.0	70.3
	PEAR$_{\text{both}}$	560M	×	81.2	59.0	70.1

Table 4: WMT24 MQM meta-evaluation. SPA and acc^*_{eq} are averaged over En-De, En-Es, and Ja-Zh; Avg Corr is the mean of these two averages. Bold indicates the best score in each column. ‡ denotes an ensemble that combines two 10.7B checkpoints and one 3.5B checkpoint. For PEAR, *ref* (†) denotes the reference-anchored configuration, and *both* denotes the bidirectional pairwise QE configuration, computed by averaging the two relative scores from both input orders. KD denotes PEAR models fine-tuned with additional MQM supervision distilled from GPT-4.1-mini.

B.3.2 Reference-Based Metrics

To provide an additional performance reference from strong reference-based metrics, we include COMET (Rei et al., 2020), BLEURT-20 (Sellam et al., 2020b), and MetricX-24-Hybrid-Large (Juraska et al., 2024). These models score single hypotheses in a reference-based setting and serve as competitive anchors alongside the QE-focused comparisons above.

C Correlation Matrices

This appendix section provides the correlation matrices for the WMT24 MQM language pairs En-De (Figure 3), En-Es (Figure 4), and Ja-Zh (Figure 5), which are discussed in Section 5.4.

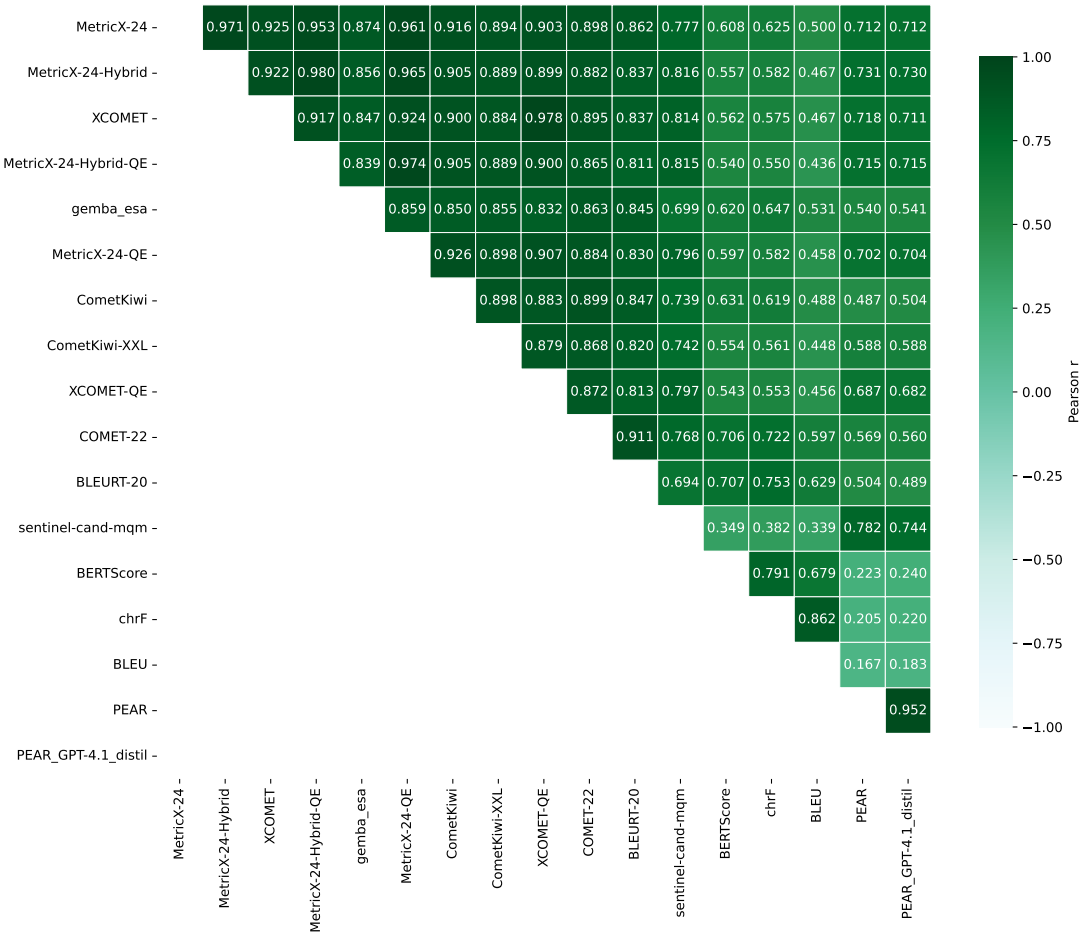

Figure 3: En-De Pearson correlation matrix between segment-level pairwise difference scores produced by WMT24 metrics. Single-candidate metrics are converted into pairwise differences by subtraction, while PEAR produces pairwise scores directly. In this matrix, PEAR corresponds to $\text{PEAR}_{\text{both}}$ and PEAR_GPT-4.1_distil corresponds to $\text{PEAR}_{\text{both,KD}}$.

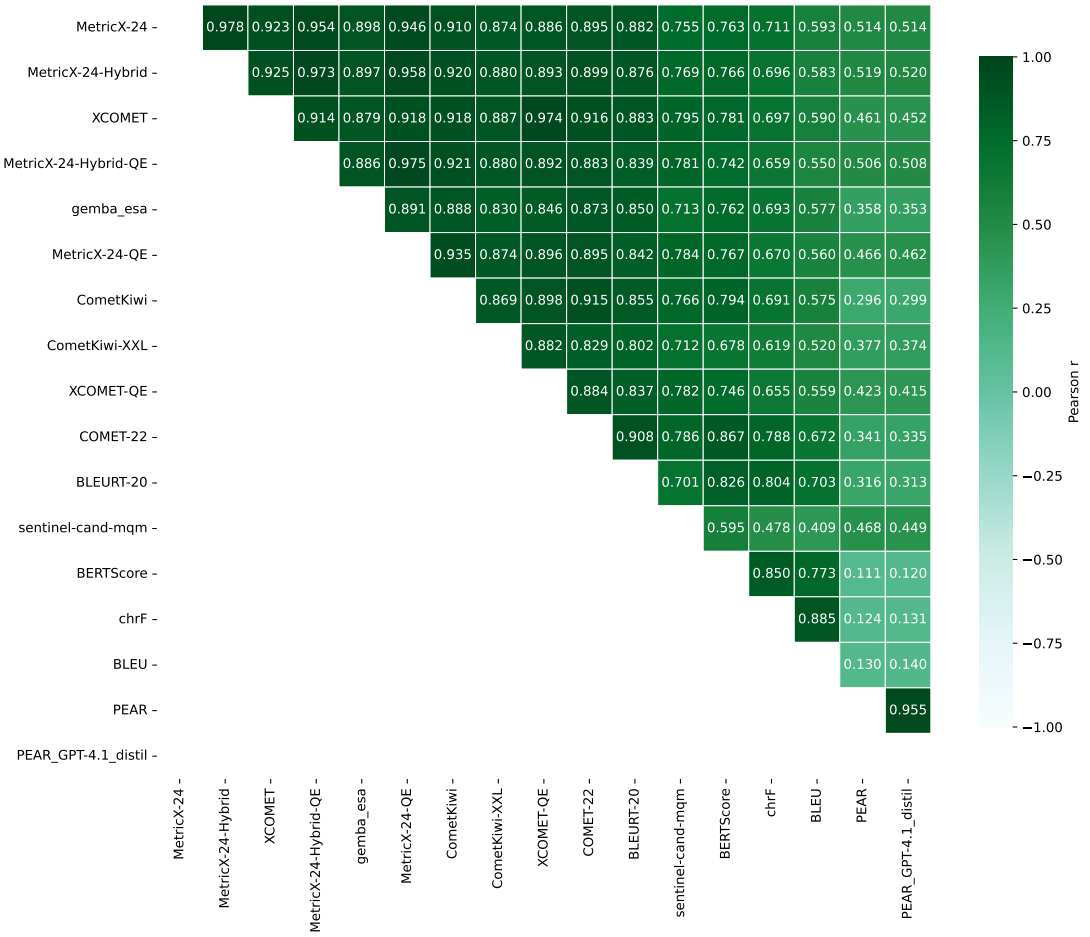

Figure 4: En-Es Pearson correlation matrix between segment-level pairwise difference scores produced by WMT24 metrics. Single-candidate metrics are converted into pairwise differences by subtraction, while PEAR produces pairwise scores directly. In this matrix, PEAR corresponds to $\text{PEAR}_{\text{both}}$ and PEAR_GPT-4.1_distil corresponds to $\text{PEAR}_{\text{both,KD}}$.

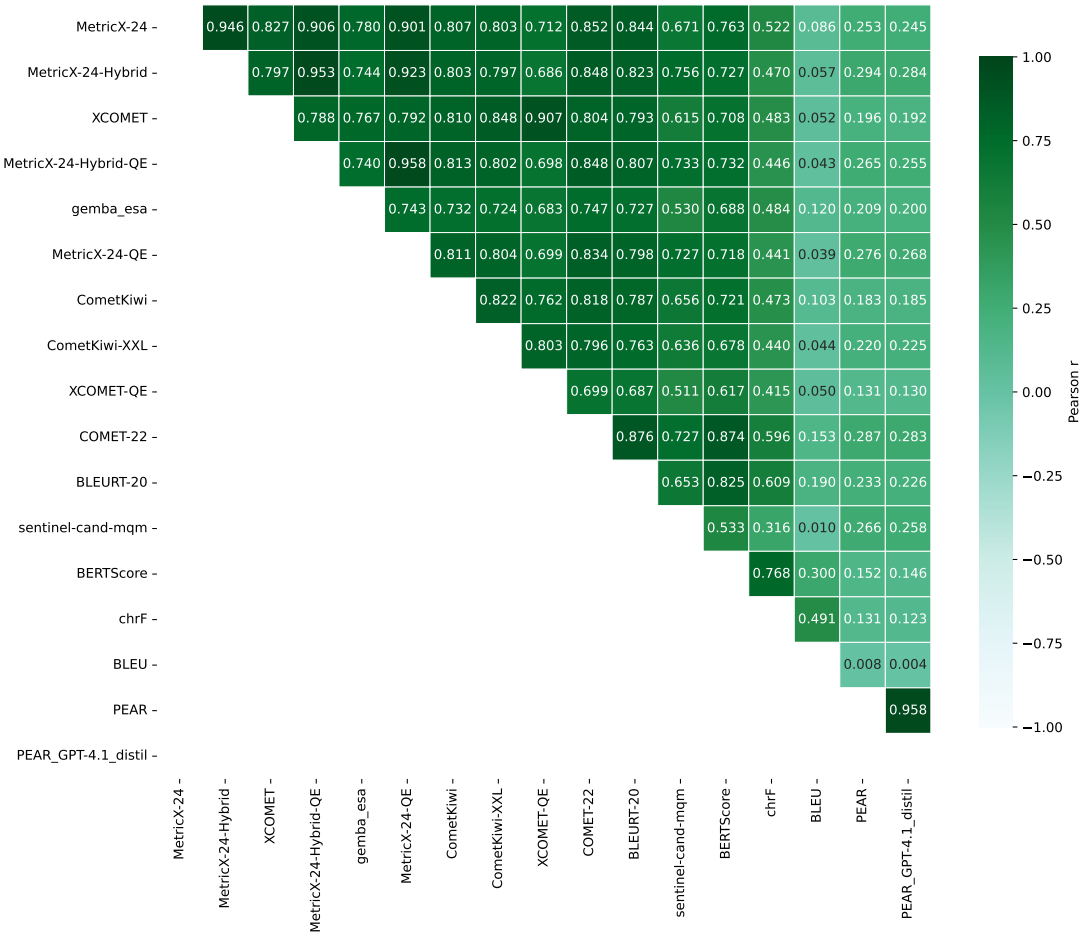

Figure 5: Ja-Zh Pearson correlation matrix between segment-level pairwise difference scores produced by WMT24 metrics. Single-candidate metrics are converted into pairwise differences by subtraction, while PEAR produces pairwise scores directly. In this matrix, PEAR corresponds to $\text{PEAR}_{\text{both}}$ and PEAR_GPT-4.1_distil corresponds to $\text{PEAR}_{\text{both,KD}}$.