Evaluating Language Translation Models by Playing Telephone

Syeda Jannatus Saba **Steven Skiena**
Stony Brook University
{sysaba, skiena}@cs.stonybrook.edu

Abstract

Our ability to efficiently and accurately evaluate the quality of machine translation systems has been outrun by the effectiveness of current language models—which limits the potential for further improving these models on more challenging tasks like long-form and literary translation. We propose an unsupervised method to generate training data for translation evaluation over different document lengths and application domains by repeated rounds of translation between source and target languages. We evaluate evaluation systems trained on texts mechanically generated using both model rotation and language translation approaches, demonstrating improved performance over a popular translation evaluation system (xCOMET) on two different tasks: (i) scoring the quality of a given translation against a human reference and (ii) selecting which of two translations is generationally closer to an original source document.

1 Introduction

Machine translation (MT) has become a fundamental pillar of multilingual communication, enabling rapid and scalable information transfer across languages. However, our ability to efficiently and accurately evaluate the quality of machine translation systems has been outrun by the effectiveness of current language models—which limits the potential for further improving these models on more challenging tasks like long-form and literary translation. Accurate translation evaluation is needed to inform model development, assess training strategies, and help with real-world deployment decisions in commercial applications.

Properly evaluating translation systems is a complex task, balancing issues of correctness and readability. Multiple valid translations exist for every given text, and proper assessment requires capturing nuances in semantics, style, and fluency. Standard metrics like BLEU (Papineni et al., 2002) rely heavily on n-gram overlap with a single human reference translation, often penalizing semantically correct yet stylistically or structurally divergent translations. Recent neural metrics such as COMET (Rei et al., 2020) and xCOMET (Guerreiro et al., 2024) use machine learning to train high-quality evaluators from human-annotated reference translations. However, human annotation is costly, and requires specialized expertise to prepare and score gold standard translations, particularly between pairs of low resource languages. These evaluation techniques will not scale as the research frontier in translation advances from single-sentence translation to the greater challenges of technical and literary translation of poetry, full-length narratives, and textbooks.

In this work, we propose a novel unsupervised method to generate training data for translation evaluation over different document lengths and application domains—without requiring any human annotation. Our approach is inspired by the popular children's game "Telephone". Here the children playing order themselves in a line, and the first child makes up a message they whisper to the second child. This process repeats until the end of the line, and with each stage in transmission some fidelity is usually lost. Merriment ensues when the last child reveals what they think the initial message was.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 10333–10348
November 4-9, 2025

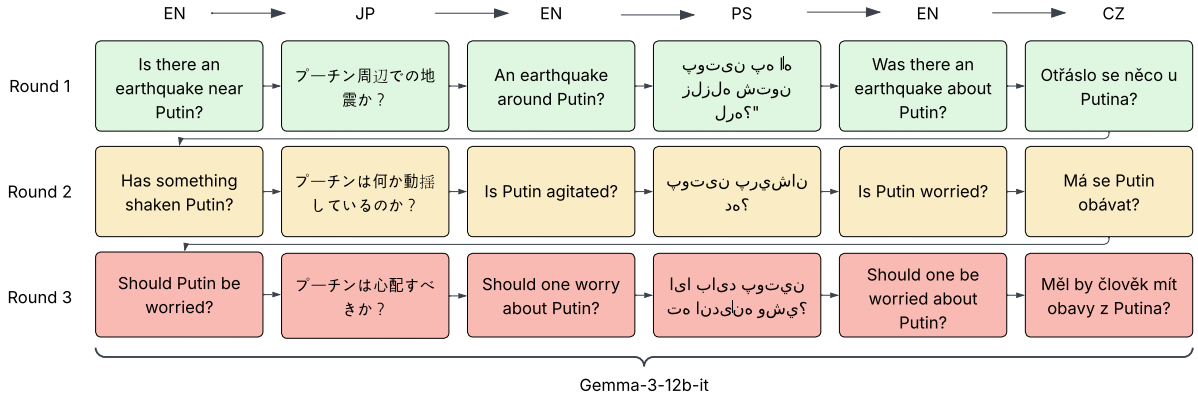


Figure 1: Language rotation setup for iterative translation. Starting from the Czech sentence “Zemětřesení kolem Putina?”, we rotate through Japanese and Pashto. Meaning gradually shifts from a factual query to subjective concern.

We leverage this idea to generate training data for language translation. Starting from an initial text, we use existing language translation models through repeated rounds of translation between source and target languages. Just as with the children’s game, each successive translation can be assumed to be more corrupted with respect to the original source than its predecessor, providing a natural measure of translation quality without human annotation. Careful analysis of such translation sequences allows us to produce pseudo-labels that capture both variations in intrinsic sentence difficulty and the dynamics of semantic drift over multiple translation cycles.

Our primary contributions include:

- *Translation evaluation without human annotation*: We introduce our “Game of Telephone” approach, providing an unsupervised pipeline to generate unlimited training data from iterative translations scored by existing automatic metrics. This method substantially reduces the barrier to developing robust MT evaluation models, especially for low-resource language scenarios where human annotations are scarce or unavailable.

  In the reference evaluation task (given a source text in language L_1 and its reference translation in another language L_2, score a second translation into L_2), our telephone-trained model outperformed (0.604) the state-of-the-art `xCOMET` (Guerreiro et al., 2024) in Pearson correlation (0.514) against the human reference standard.

- *Language rotation vs. model rotation*: We systematically explore two variants of our approach: language-rotated translation cycles (varying the pivot languages at each iteration) and model-rotated translation cycles (varying MT models at each iteration). We empirically demonstrate both strategies can capture nuanced translation degradation: both language- and model-rotation trained models obtained higher AUC scores than `xCOMET` at predicting the earlier generation texts for five different LLMs.

- *New translation evaluation benchmark*: We have released evaluation metrics (both code and data) to assess translator systems for sensitivity to semantic drift as well as sentence-level accuracy/robustness[1]. This serves as a resource for advancing language translation quality assessment methodologies.

Our paper is organized as follows. We survey previous work in translation evaluation in Section 2. Details of our approach to quality scoring as a function of translation round, document difficulty, and relative model fidelity are discussed in Section 3. We explain how we use this data to train evaluation models in Section 4, with our experimental

[1]https://github.com/saba-phoenix/language-translation-evaluation-by-playing-telephone

results in Section 5. We conclude with ideas for future research in Section 6.

2 Previous Work

Machine Translation. MT has advanced with large transformers—from the original Transformer (Vaswani et al., 2017) to multilingual models like mBART (Liu et al., 2020), M2M-100 (Fan et al., 2021), and NLLB-200 (Costa-jussà et al., 2022). Instruction-tuned LLMs such as GPT-4 (OpenAI, 2023), Mixtral (Jiang et al., 2024), and Gemini (Team et al., 2023) now handle translation as a byproduct of language understanding. Still, low-resource performance lags (Ruder et al., 2021; Freitag et al., 2022a), and issues like idioms and domain shift demand better evaluation and adaptation.

Translation Evaluation System. While MT models have advanced rapidly, evaluation metrics have lagged behind. Traditional metrics such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and TER (Snover et al., 2006) rely on surface-level n-gram overlap, frequently failing to capture semantic adequacy or fluency. This gap has driven the development of learned metrics more aligned with human judgment, such as `COMET` and `BLEURT` (Sellam et al., 2020). Benchmarks like WMT (Freitag et al., 2021b), MLQE-PE (Fomicheva et al., 2020), and FLORES (Team et al., 2022) have become central to evaluating translation quality.

COMET and xCOMET. Learned metrics like `COMET` and its extensions `COMETKIWI` (Rei et al., 2022), `xCOMET` have improved MT evaluation by aligning closely with human judgments, incorporating uncertainty estimates (Glushkova et al., 2021) and span-level error detection. Yet recent findings (Agrawal et al., 2023) reveal a saturation effect at the upper end of the quality spectrum, where these metrics struggle to distinguish between strong translations. This is particularly limiting when comparing top-tier systems or reranking hypotheses.

Synthetic Data Generation. Tuan et al. (2021) create synthetic QE data by applying MT and MLM rewriting over mined parallel corpora. The main idea is to inject noise into clean translations, and then recover word-level tags through edit distance with a pseudo-reference. Etchegoyhen and Ponce (2023) collect translations from MT models at different training checkpoints, using earlier checkpoints to simulate weaker systems and pairing them with final model outputs as references.

Translation Chains and Semantic Drift. Beyond static evaluations, real-world applications often involve translation chains, such as multilingual pivots or model rotations, where quality can drift gradually. Prior work on cyclic translation obfuscation (Potthast et al., 2013) formalizes this idea, using multi-hop translations across diverse languages and engines to paraphrase content while preserving meaning. Since different intermediate languages reorder, compress, or expand information differently (Beinborn and Choenni, 2020), such paths naturally introduce subtle distortions. Existing metrics overlook these dynamics. To address these gaps, we construct a degradation-aware scoring framework built on iterative translations. We now describe how we transform raw metric outputs into more meaningful learning signals.

3 Scoring Telephone Translations

In our Telephone setup, a sentence is translated through multiple rounds of MT, each introducing potential semantic drift through model- or language-rotation. This setting exposes two hidden factors that influence translation quality:

- **Global sentence difficulty:** Some sentences degrade quickly regardless of the iteration or system used.
- **Local iteration pressure:** Some iterations are inherently harder due to the model or pivot language used.

Standard metrics like `xCOMET` overlook these, treating outputs as independent and context-free. We transform raw metric scores into a more informative, context-aware scoring signal by accounting for sentence fragility and iteration difficulty, yielding a refined score that tracks translation quality without human supervision.

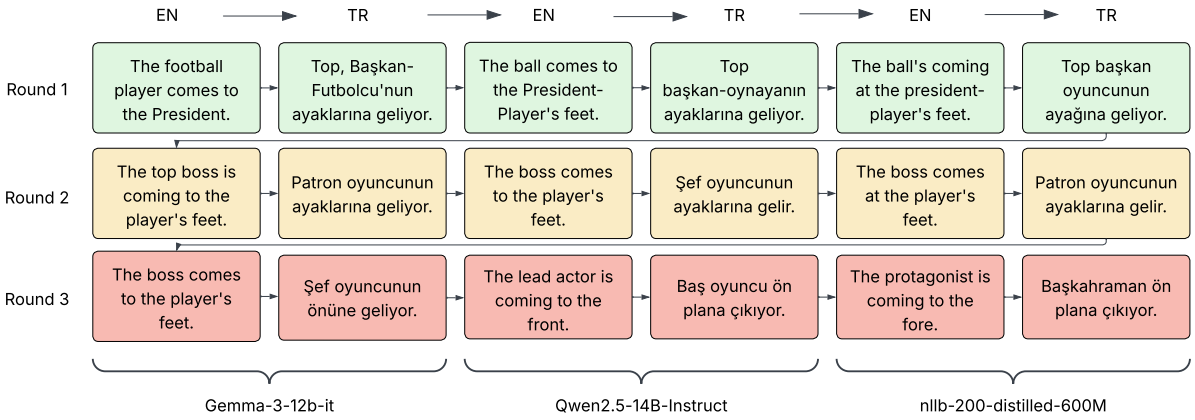


Figure 2: Model rotation for iterative translation. We perform three rounds of Turkish-English round-trip translations on "Futbolcu Başkan'ın ayağına gelir". Later iterations reveal compounding semantic drift.

3.1 Score Normalization Procedure

Formally, given a source sentence s_i from a dataset of N sentences, and its translation $T_i^{(j)}$ at iteration j, we compute raw scores $q_i^{(j)}$ using an automatic metric (xCOMET), and aim to transform them into context-aware scores $r_i^{(j)}$.

Step 1: Estimate Sentence Fragility. To capture sentence-specific intrinsic translation difficulty, we define a sentence-level hardness measure by averaging each sentence's quality scores over all K iterations and standardize it across the dataset:

$$z_i = \frac{\mu_i - \bar{\mu}}{\bar{\sigma}} \tag{1}$$

where $\bar{\mu}$ and $\bar{\sigma}$ are the mean and standard deviation of μ_i across the dataset. This z_i serves as a global indicator of how inherently difficult sentence s_i is to translate.

Step 2: Estimate Iteration Pressure. We characterize each iteration's overall translation quality profile by computing iteration-specific statistics, namely the mean $\mu^{(j)}$ and standard deviation $\sigma^{(j)}$ across the raw scores of all sentences at iteration j.

$$\mu^{(j)} = \frac{1}{N}\sum_{i=1}^{N} q_i^{(j)} \tag{2}$$

$$\sigma^{(j)} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(q_i^{(j)} - \mu^{(j)})^2} \tag{3}$$

where N is the number of sentences.

Step 3: Compute Refined Score. Finally, we produce the refined, context-aware translation score $r_i^{(j)}$ by re-projecting each sentence's global difficulty z_i into the local score distribution at iteration j:

$$r_i^{(j)} = \mu^{(j)} + z_i \cdot \sigma^{(j)} \tag{4}$$

This cross-distribution projection yields a refined score $r_i^{(j)}$ that is designed to capture how difficult sentence s_i is to translate under the pressure of iteration j. The transformation helps disentangle raw metric scores from local and global confounders, producing a degradation-aware signal suitable for evaluator training or analysis.

4 Training Evaluation Models

4.1 Training Data Construction

We use 114,864 source-reference pairs from WMT DA (2018–2022) (Kocmi et al., 2022), across 36 language pairs, and split them 80/20 into training and validation. We build iterative translation chains where each sentence undergoes 18 translation rounds. This number arises from our design: 3 rotation setups (model or language triplets) $\times$ 2 directions of translation per iteration (forward and back) $\times$ 3 iterations per setup = 18. The choice ensures (i) $\sim$1M synthetic examples, comparable to the 1M DA-labeled examples used in xCOMET pretraining, and (ii) at least three complete forward iterations per setup, which are required for our paired-generation experiments in Subsection 5.2.

We consider two complementary rotation strategies for constructing these chains:

Model Rotation. In this setup (Figure 2), we fix the language pair and rotate the MT model at each iteration. Each sentence is translated and back-translated using one model, then passed to the next in a cyclic order. This avoids convergence artifacts like idempotency and introduces diverse distortions that accelerate semantic drift. As shown in Figure 3, model rotation leads to faster degradation compared to static systems.

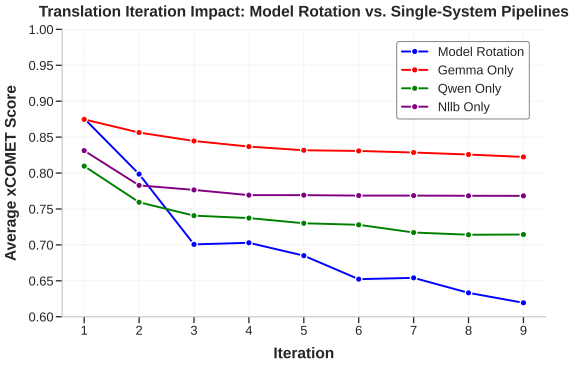


Figure 3: Average `xCOMET` scores over forward 9 iterations for 300 sentences. Model rotation degrades fastest; single-model pipelines plateau after early decline.

Our training rotation involves three distinct MT systems: `gemma-3-12b-it` (Anil et al., 2024), `Qwen2.5-14B-Instruct` (Cui et al., 2024), and `nllb-200-distilled-600M` (Costa-jussà et al., 2022).

Language Rotation. In this setup (Figure 1), we fix the MT model and rotate the target language at each iteration. Each sentence starts in a source language, is translated into English, then into another language, continuing cyclically through a triplet of languages (not including English). We experiment with two types of triplets: low-diversity (genealogically and geographically similar languages) and high-diversity (genealogically and geographically distant, often including low-resource languages). In a variant of this setup, we bypass English entirely and rotate translations directly among the three languages without a fixed pivot.

Regarding the selection of intermediary languages for our language rotation strategy: we started with a pool of 24 languages and used GPT-4o (OpenAI, 2023) to generate initial candidate clusters based on general linguistic similarity. These initial groupings were then manually refined by the authors using typological and genealogical resources: WALS (Dryer and Haspelmath, 2011) and Glottolog (Hammarström et al., 2023). The final sets of triplets are provided in Appendix Tables 6 and 7.

We group translation chains that share the same model and language setup, compute their refined scores, and use those scores as training supervision for our evaluation models.

4.2 Model Architecture and Training

We adopt the `COMET` and `xCOMET` architecture (Rei et al., 2020, 2022; Guerreiro et al., 2024) and train models under three standard input modes: reference-free or Quality Estimation (SRC), reference-based Regression (SRC+REF), and unified (SRC, REF, MT). The reference-free or QE mode predicts quality from source and translation only; the unified mode (Wan et al., 2022) supports all input combinations. Inputs are constructed by embedding the source, machine translation, and/or reference (depending on the configuration), and combining their embeddings. All inputs are passed through a pretrained multilingual encoder: `XLM-RoBERTa-large` (Conneau et al., 2020) for the QE and regression models, and `InfoXLM-Large` (Chi et al., 2021) for the unified variant. Layer representations from the encoder are combined using a sparsemax-weighted scalar mix (Martins and Astudillo, 2016). The resulting [CLS] token is fed into a two-layer feedforward network to produce the final sentence-level quality score.

All models are trained using mean squared error (MSE) against our degradation-aware refined scores. We freeze encoder embeddings for the first 30% of training and use the AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of 1.5e-5 (1e-6 for the encoder), batch sizes of 16–32, and layerwise learning rate decay of 0.95. Training runs for up to 5 epochs with early stopping on validation loss.

4.3 Model Variants

We present all model variants, ours and existing automatic metrics that we use as a baseline; grouped by reference usage and supervision type, as summarized in Table 1.

Alongside human-provided references, we also experiment with pseudo-references: translations from the previous iterations in a degradation chain. Models marked "REL" use pseudo-references, while "HM" models use gold references from WMT DA.

To further improve evaluation quality, we finetuned our best-performing models (UM-MR-HM and QE-MR) on human-annotated MQM data (Freitag et al., 2021a,c, 2022b), producing the UM-FT-MQM and QE-FT-MQM variants. This two-stage training strategy: pretraining followed by MQM finetuning, follows the approach introduced in xCOMET (Rei et al., 2023). The MQM dataset includes 148k unique examples across zh-en, en-de, and en-ru, split into training (104.8k), validation (14.5k), and test (28.9k) sets. The test set is later used to assess correlation with human judgments.

Model Name	Description
Baseline QE Metrics	
xCOMET-NR	xCOMET without reference input.
COMETKIWI	QE-style COMET trained on WMT DA scores.
Our QE Metrics	
QE-MR	QE model trained on model-rotated degradation data.
QE-FT-MQM	QE-MR fine-tuned on MQM human annotations.
QE-LR-LD	QE model+low-diversity language rotation (Eng. pivot).
QE-LR-HD	QE model+ high-diversity language rotation (Eng. pivot).
QE-LR-LDD	QE model+low-diversity direct multilingual transitions.
QE-LR-HDD	QE model+high-diversity direct multilingual transitions.
Baseline Reference-Based and Unified Metrics	
xCOMET	COMET trained with human references.
COMETDA	COMET trained directly on WMT DA scores.
Our Reference-Based and Unified Metrics	
UM-MR-HM	Unified Metric+human reference (model-rotation).
UM-MR-REL	Unified Metric+pseudo-reference (prior iteration).
UM-FT-MQM	UM-MR-HM fine-tuned on MQM human annotations.
REG-MR-HM	Regression model+human reference (model-rotation).
REG-MR-REL	Regression model+pseudo-reference.

Table 1: Summary of model variants

These variants allow us to investigate the trade-offs between supervised, pseudo-supervised, and unsupervised evaluator training, as well as the efficacy of degradation-based supervision across different configurations.

Metric	en-de	en-ru	zh-en	Average
QE Metrics				
xCOMET-NR	0.481	0.412	0.531	0.475
COMETKIWI	0.371	0.339	0.335	0.348
QE-LR-LD	0.346	0.228	0.495	0.356
QE-LR-HD	0.331	0.213	0.493	0.346
QE-LR-LDD	0.375	0.275	0.474	0.375
QE-LR-HDD	0.316	0.203	0.456	0.325
QE-MR	0.378	0.297	0.513	0.396
QE-FT-MQM	**0.485**	**0.533**	**0.689**	**0.569**
Reference-Based and Unified Metrics				
xCOMET	**0.532**	0.455	0.556	0.514
COMETDA	0.434	0.402	0.394	0.410
REG-MR-HM	0.467	0.371	0.532	0.457
REG-MR-REL	0.440	0.306	0.519	0.422
UM-MR-HM	0.488	0.414	0.551	0.484
UM-MR-REL	0.478	0.389	0.552	0.473
UM-FT-MQM	0.523	**0.696**	**0.592**	**0.604**

Table 2: Pearson correlation with human annotated MQM scores across three language pairs for QE, reference-based, and unified metrics.

5 Experimental Results

5.1 Human Annotation Evaluation

We compare our telephone-supervised evaluation models to existing metrics using WMT MQM human annotations. Table 2 reports segment level Pearson correlations with human scores; results for other correlation metrics and DA annotations are provided in Appendix Tables 8–10.

To strengthen our segment level evaluation setup, we follow the WMT24 shared task design (Freitag et al., 2024) and complement correlation with "group-by-item" accuracy with tie calibration acc_{eq} (Deutsch et al., 2023), which measures agreement with human segment-level preferences in pairwise comparisons.

At the system level, we adopt Soft Pairwise Accuracy (SPA) (Thompson et al., 2024). Unlike standard pairwise accuracy or Kendall's τ, SPA incorporates the statistical uncertainty of both metric and human rankings, ensuring that near-ties are handled consistently rather than penalized or rewarded arbitrarily.

Pearson correlation is reported over the full MQM test set. In contrast, acc_{eq} and SPA require source sentences to be available across all compared systems; for these evaluations, we selected

Metric	en-de		en-ru		zh-en		Average	
	acc_{eq}	SPA	acc_{eq}	SPA	acc_{eq}	SPA	acc_{eq}	SPA
QE Metrics								
xCOMET-NR	0.454	**0.932**	0.519	0.763	0.449	0.561	0.474	0.752
COMETKIWI	0.383	0.901	0.504	0.925	0.462	0.536	0.450	0.787
QE-MR	0.475	0.713	0.481	0.785	0.454	0.804	0.470	0.767
QE-FT-MQM	**0.508**	0.919	**0.659**	**0.931**	0.452	0.635	**0.540**	**0.828**
QE-LR-LD	0.432	0.683	0.450	0.721	**0.491**	0.563	0.458	0.656
QE-LR-HD	0.448	0.700	0.426	0.764	0.446	**0.812**	0.440	0.759
QE-LR-LDD	0.426	0.790	0.465	0.771	0.441	0.441	0.444	0.667
QE-LR-HDD	0.475	0.746	0.411	0.770	0.433	0.708	0.440	0.741
Reference-Based and Unified Metrics								
xCOMET	0.552	0.905	0.597	0.728	0.462	0.712	**0.537**	0.782
COMETDA	**0.585**	**0.975**	0.488	**0.957**	0.452	0.711	0.508	**0.881**
REG-MR-HM	0.503	0.906	0.488	0.867	0.462	**0.827**	0.484	0.867
REG-MR-REL	0.541	0.736	0.450	0.801	0.472	0.762	0.488	0.766
UM-MR-HM	0.492	0.940	0.566	0.844	**0.483**	0.531	0.514	0.772
UM-MR-REL	0.497	0.925	0.597	0.856	0.470	0.429	0.521	0.737
UM-FT-MQM	0.497	0.902	**0.605**	0.812	0.464	0.630	0.522	0.781

Table 3: Segment and system-level meta-evaluation using acc_{eq} and SPA across three language pairs.

the top three systems per language pair and used only the subset of sources common to them.

5.1.1 Segment-Level Meta-Evaluation

Reference-Free Metrics. From Pearson correlation data on Table 2, model rotation proves the most effective degradation strategy: QE-MR (0.396) outperforms all language-rotated variants, with lower-diversity setups (e.g., QE-LR-LDD at 0.375) outperforms high-diversity ones, hinting that excessive linguistic variation may dilute the training signal. Despite being trained entirely without human labels, QE-MR performs competitively and surpasses COMETKIWI (0.348). Fine-tuning on MQM further boosts performance: QE-FT-MQM leads all reference-free models with 0.569, outperforming xCOMET-NR (0.475). These results suggest that degradation-based pretraining provides a solid foundation, strong on its own, and even stronger when combined with human-labeled supervision. The same pattern is consistently reflected under acc_{eq} results in Table 3.

Reference-Based Metrics. UM-FT-MQM leads with the highest average Pearson correlation (0.604), showing strong gains on en–ru (0.696) and zh–en (0.592), and surpassing the established xCOMET (0.514). Our degradation-trained models, especially UM-MR-HM (0.484) and UM-MR-REL (0.473), closely trail behind, showing that pseudo-references can be nearly as effective as gold ones. Regression variants like REG-MR-HM (0.457) and REG-MR-REL (0.422) remain competitive, while COMETDA (0.410) trails slightly, likely due to its training with DA-specific data.

When evaluated with acc_{eq} (Table 3), xCOMET (0.537) achieves the strongest agreement with human segment-level preferences, followed closely by UM-FT-MQM (0.522) and UM-MR-REL (0.521). UM-MR-HM (0.514) remains competitive. Interestingly, COMETDA, which lagged on Pearson correlation, performs reasonably well under acc_{eq} (0.508), indicating that while it struggles with fine-grained score alignment, it remains effective at identifying the better translation in pairwise comparisons.

Category	System	1 vs 2			2 vs 3			1 vs 3		
		xCOMET	LR	MR	xCOMET	LR	MR	xCOMET	LR	MR
Model Rotation	Gemma-12b	0.876	0.935	**0.937**	0.663	**0.718**	0.712	0.907	0.957	**0.958**
	Qwen-2.5-14b	0.791	0.850	**0.864**	0.640	0.681	**0.682**	0.832	0.888	**0.902**
	NLLB-distilled	0.677	**0.753**	0.745	0.599	**0.644**	0.635	0.723	**0.806**	0.798
Commercial Models MR	GPT-4o	0.955	**0.973**	**0.973**	0.628	**0.706**	0.669	0.966	**0.983**	0.981
	GPT-4o-mini	0.942	**0.965**	0.963	0.641	**0.703**	0.678	0.958	**0.976**	0.974
	NLLB-distilled	0.879	**0.931**	0.917	0.616	**0.685**	0.649	0.905	**0.950**	0.938
Language Rotation	First Language	0.732	**0.865**	0.830	0.737	**0.847**	0.797	0.852	**0.955**	0.927
	Second Language	0.727	**0.839**	0.783	0.688	**0.774**	0.727	0.820	**0.916**	0.872
	Third Language	0.759	**0.845**	0.801	0.698	**0.769**	0.738	0.839	**0.913**	0.885

Table 4: AUC scores for detecting translation quality degradation across three round pairs (1 vs 2, 2 vs 3, 1 vs 3), comparing xCOMET, language-rotation (LR), and model-rotation (MR) variants across different setups. Higher AUC indicates better discrimination between earlier and later degraded outputs.

5.1.2 System-Level Meta-Evaluation

Reference-Free Metrics. SPA results in Table 3 reveal that reference-free models generally lag behind their reference-based counterparts in system-level ranking. Among them, QE-FT-MQM stands out with the highest SPA (0.828), showing that fine-tuning on MQM not only improves segment-level correlation but also enhances robustness when comparing systems. Interestingly, high diversity language-rotation variants such as QE-LR-HD (0.759) and QE-LR-HDD (0.741) achieve relatively strong SPA despite weaker segment level scores.

Reference-Based Metrics. Here SPA highlights a different set of winners than Pearson correlation. The baseline COMETDA achieves the highest score (0.881), while xCOMET reaches 0.782. Among our models, regression variants perform best, with REG-MR-HM at 0.867. Unified metrics such as UM-FT-MQM (0.781) and UM-MR-HM (0.772) remain competitive but do not surpass the baselines, reflecting their strength in segment-level evaluation rather than system-level ranking.

5.2 Paired Generation Evaluation

A core assumption of our framework is that earlier iterations in a translation chain should retain higher quality than later ones. To test whether evaluation metrics capture this implicit ordering, we conduct a paired generation test: given two outputs (e.g., from round 1 and 3), can the metric correctly identify the better one? We treat this as a binary discrimination task and report AUC scores for all pairwise comparisons among rounds: 1 vs 2, 2 vs 3, and 1 vs 3, using outputs from model-rotated chains, language-rotated chains, and model-rotated chains including commercial systems like GPT-4o. We compare three representative metrics: xCOMET, our language-rotation model QE-LR-LD (LR), and its model-rotation counterpart QE-MR (MR).

As shown in Table 4, distinguishing between later iterations, especially 2 vs 3 is consistently the most difficult. AUC scores drop sharply here, with xCOMET reaching just 0.599 on NLLB and 0.628 on GPT-4o. In contrast, our degradation-trained models show greater resilience: MR improves to 0.635 and 0.669 on the same systems, while LR does even better, achieving 0.644 and 0.706.

Surprisingly, LR despite scoring lower on human correlation, outperforms MR in 8 out of 9 commercial comparisons and performs comparably in model-rotation settings. Its advantage is most pronounced in language-rotated chains, where it achieves up to 0.847 (2 vs 3) and 0.955 (1 vs 3), suggesting a stronger capacity to track structural shifts across languages. On high-quality outputs from GPT-4o, both LR and MR maintain near-ceiling AUCs (up to 0.983), consistently outperforming xCOMET across all comparisons.

Training Signal Variants. We compare three training strategies on the paired generation test (Table 5). UM-MR-HM uses our refined scores. IT ignores sentence-level detail, assigning a single aver-

Model	1 vs 2			2 vs 3			1 vs 3		
	UM	IT	UNMD	UM	IT	UNMD	UM	IT	UNMD
Gemma-12b	0.944	**0.989**	0.913	0.754	**0.789**	0.692	0.962	**0.993**	0.937
Qwen-2.5-14b	0.882	**0.943**	0.825	0.719	**0.739**	0.665	0.913	**0.961**	0.864
NLLB-distilled	0.786	**0.834**	0.710	0.675	**0.679**	0.614	0.835	**0.883**	0.757

Table 5: AUC scores for degradation detection across round pairs (1 vs 2, 2 vs 3, 1 vs 3), comparing UM-MR-HM (UM), iteration-averaged (IT), and unmodified xCOMET-trained (UNMD) models on three MT systems.

age score per iteration. UNMD is trained directly on xCOMET scores, without degradation awareness. IT performs best across all systems; for example, scoring 0.961 (1 vs 3) and 0.739 (2 vs 3) on Qwen-2.5, outperforming UM-MR-HM at 0.913 and 0.719. Since IT only knows which iteration a sentence came from, it learns to align position with quality, an advantage in relative ranking. UM-MR-HM still performs well, showing that sentence-level signals help capture more detail even if they aren't essential for ranking which output is better. UNMD lags behind, confirming the value of degradation-based supervision.

6 Conclusion

We have demonstrated that synthetic translation data produced using language and model rotation strategies can be used to train state-of-the-art translation evaluation models. Future work should focus on better methods to score the training data as a function of generation and text complexity, and applying this methodology to more challenging long-form tasks.

Limitations

While promising, our method has several limitations. First, the degradation trajectories we induce may not fully reflect real-world translation errors. They tend to emphasize lexical and syntactic shifts, potentially underrepresenting issues such as discourse inconsistency or cultural misalignment. Additionally, the quality of our pseudo-labels ultimately depends on the reliability of the scoring anchor (xCOMET); any bias or blind spots in this metric propagate to the training signal. Finally, we assume access to a diverse set of MT systems and high-coverage multilingual translation capabilities. In truly low-resource settings with poor model availability, both rotation strategies may be difficult to execute meaningfully, constraining the generality of our approach.

References

Sweta Agrawal, Antônio Farinhas, Ricardo Rei, and André F. T. Martins. 2023. Can automatic metrics assess high-quality translations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics.

Rohan Anil, Yanping Huang, Andrew M Dai, Sharan Narang, Yifeng Lu, Aakanksha Chowdhery, David Dohan, Aditya Siddhant, Daniel De Freitas, Anselm Levskaya, and et al. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.07672.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL Workshop*, pages 65–72.

Lisa Beinborn and Rochelle Choenni. 2020. Semantic drift in multilingual representations. *Computational Linguistics*, 46(3):571–603.

Zewen Chi, Li Dong, Furu Wei, Wenhui Wang, 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. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard 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. Association for Computational Linguistics.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Yuxiao Cui, Qingyang Bai, Weizhi Wang, Junyang Lin, Bohan Zhuang, Lei Li, and et al. 2024. Qwen2: Scaling up language models in chinese and english. *Preprint*, arXiv:2401.04088.

Daniel Deutsch, George Foster, and Markus Freitag. 2023. Ties matter: Meta-evaluating modern metrics with pairwise accuracy and tie calibration. *arXiv preprint arXiv:2305.14324*.

Matthew S. Dryer and Martin Haspelmath, editors. 2011. *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Thierry Etchegoyhen and David Ponce. 2023. Learning from past mistakes: Quality estimation from monolingual corpora and machine translation learning stages. In *Proceedings of Machine Translation Summit XIX, Vol 1: Research Track*, pages 84–98, Macau, China. Asia-Pacific Association for Machine Translation.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22:1–48.

Marina Fomicheva, Dimitar Shterionov, Francisco Guzmán, Lucia Specia, and André FT Martins. 2020. Unsupervised quality estimation for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Markus Freitag, Yaser Al-Onaizan, Shuo Sun Ma, and 1 others. 2022a. High-quality low-resource machine translation: A new benchmark. In *Findings of EMNLP*.

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, 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, Prashant Mathur, Ondřej Bojar, and 1 others. 2021b. Results of the wmt21 metrics shared task: Evaluating metrics with expert-based human evaluations on ted and news tasks. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. 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*, pages 46–68, Abu Dhabi. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021c. Results of the wmt21 metrics shared task: Evaluating metrics with expert-based human evaluations on ted and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

Taisiya Glushkova, Ricardo Rei, Ana Farinha, and Lucia Specia. 2021. Uncertainty-aware comet: A confidence estimation framework for mt evaluation. In *WMT*, pages 1026–1035.

Nuno M Guerreiro, Ricardo Rei, Daan van Stigt, Luisa 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.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2023. Glottolog 4.8.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

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, Maja Popović, and Mariya Shmatova. 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.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*. OpenReview preprint.

André F. T. Martins and Ramon Astudillo. 2016. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, volume 48, pages 1614–1623, New York, NY, USA. PMLR.

OpenAI. 2023. GPT-4 technical report. `https://openai.com/research/gpt-4`.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

Martin Potthast, Alberto Barrón-Cedeño, Benno Stein, and Paolo Rosso. 2013. Overview of the 5th international competition on plagiarism detection. In *Working Notes for CLEF 2013 Conference*, volume 1179, pages 1–20. CEUR Workshop Proceedings.

Ricardo Rei, Ana Farinha, Alon Lavie, and Lucia Specia. 2020. Comet: A neural framework for mt evaluation. In *EMNLP*, pages 2685–2702.

Ricardo Rei, António Farinhas, André F. T. Martins, and Lucia Specia. 2023. xcomet: Transparent machine translation evaluation through fine-grained error detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8334–8352. 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)*. Association for Computational Linguistics.

Sebastian Ruder, Alessandro Raganato, Marcin Staniszewski, Shruti Singh, and et al. 2021. XTREME-r: Towards more challenging and nuanced multilingual evaluation. *arXiv preprint*, arXiv:2104.07412.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 223–231. Association for Machine Translation in the Americas.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

NLLB Team, Marta R. Costa-jussà, James Cross, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Brian Thompson, Nitika Mathur, Daniel Deutsch, and Huda Khayrallah. 2024. Improving statistical significance in human evaluation of automatic metrics via soft pairwise accuracy. *arXiv preprint arXiv:2409.09598*.

Yi-Lin Tuan, Ahmed El-Kishky, Adithya Renduchintala, Vishrav Chaudhary, Francisco Guzmán, and Lucia Specia. 2021. Quality estimation without human-labeled data. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 619–625, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008. Curran Associates, Inc.

Yu Wan, Dayiheng Liu, Baosong Yang, Haibo Zhang, Boxing Chen, Derek Wong, and Lidia Chao. 2022. UniTE: Unified translation evaluation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8117–8127, Dublin, Ireland. Association for Computational Linguistics.

A Appendix

Source Language	Low-Diversity Triplet
Bengali	Bengali, Hindi, Gujarati
Central Khmer	Central Khmer, Chinese, Japanese
Chinese	Chinese, Central Khmer, Japanese
Czech	Czech, Polish, Russian
Estonian	Estonian, Finnish, Latvian
Finnish	Finnish, Estonian, Latvian
French	French, German, Polish
German	German, French, Polish
Gujarati	Gujarati, Hindi, Bengali
Hausa	Hausa, Zulu, Xhosa
Hindi	Hindi, Gujarati, Bengali
Icelandic	Icelandic, German, French
Japanese	Japanese, Chinese, Central Khmer
Kazakh	Kazakh, Russian, Ukrainian
Latvian	Latvian, Lithuanian, Estonian
Lithuanian	Lithuanian, Latvian, Estonian
Pashto	Pashto, Hindi, Kazakh
Polish	Polish, Czech, German
Russian	Russian, Ukrainian, Kazakh
Tamil	Tamil, Hindi, Gujarati
Turkish	Turkish, Kazakh, Russian
Ukrainian	Ukrainian, Russian, Kazakh
Xhosa	Xhosa, Zulu, Hausa
Zulu	Zulu, Xhosa, Hausa

Table 6: Low-diversity triplets used in our experiments.

Source Language	High-Diversity Triplet
Bengali	Bengali, Russian, Hausa
Central Khmer	Central Khmer, Turkish, Finnish
Chinese	Chinese, Zulu, Lithuanian
Czech	Czech, Japanese, Pashto
Estonian	Estonian, Hindi, Xhosa
Finnish	Finnish, Tamil, Kazakh
French	French, Gujarati, Central Khmer
German	German, Pashto, Japanese
Gujarati	Gujarati, Ukrainian, Zulu
Hausa	Hausa, Chinese, Latvian
Hindi	Hindi, Estonian, Kazakh
Icelandic	Icelandic, Bengali, Japanese
Japanese	Japanese, Hausa, Latvian
Kazakh	Kazakh, Xhosa, Polish
Latvian	Latvian, Tamil, Japanese
Lithuanian	Lithuanian, Hindi, Zulu
Pashto	Pashto, French, Chinese
Polish	Polish, Central Khmer, Gujarati
Russian	Russian, Estonian, Zulu
Tamil	Tamil, German, Ukrainian
Turkish	Turkish, Japanese, Lithuanian
Ukrainian	Ukrainian, Hausa, Chinese
Xhosa	Xhosa, Russian, Hindi
Zulu	Zulu, French, Kazakh

Table 7: High-diversity triplets used in our experiments.

Metric	**en-de**	**en-ru**	**zh-en**	**Average**
QE Metrics				
xCOMET-NR	**0.410**	0.468	0.507	0.462
COMETKIWI	0.306	0.391	0.372	0.356
QE-LR-LD	0.315	0.312	0.472	0.366
QE-LR-HD	0.314	0.284	0.478	0.359
QE-LR-LDD	0.342	0.342	0.465	0.383
QE-LR-HDD	0.326	0.286	0.480	0.364
QE-MR	0.328	0.366	0.491	0.395
QE-FT-MQM	0.368	**0.474**	**0.558**	**0.467**
Reference-Based and Unified Metrics				
xCOMET	**0.476**	0.513	0.528	0.506
COMETDA	0.419	0.433	0.439	0.430
REG-MR-HM	0.412	0.407	0.507	0.442
REG-MR-REL	0.401	0.391	0.496	0.429
UM-MR-HM	0.436	0.442	0.519	0.466
UM-MR-REL	0.422	0.418	0.514	0.451
UM-FT-MQM	0.432	**0.581**	**0.536**	**0.516**

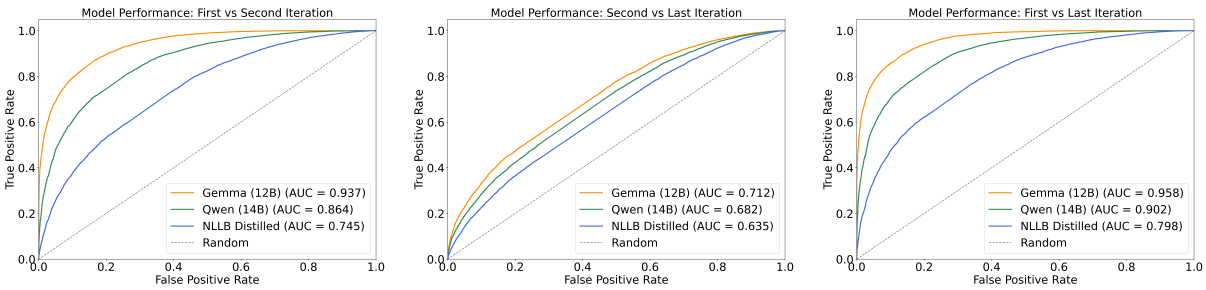
Table 8: Spearman correlation with human annotated MQM scores across three language pairs for QE, reference-based, and unified metrics.

Metric	en-de	en-ru	zh-en	Average
QE Metrics				
xCOMET-NR	**0.320**	**0.357**	0.382	0.353
COMETKIWI	0.229	0.288	0.271	0.263
QE-LR-LD	0.235	0.228	0.350	0.271
QE-LR-HD	0.234	0.206	0.353	0.264
QE-LR-LDD	0.257	0.251	0.343	0.284
QE-LR-HDD	0.243	0.208	0.354	0.268
QE-MR	0.246	0.269	0.364	0.293
QE-FT-MQM	0.284	0.354	**0.431**	**0.356**
Reference-Based and Unified Metrics				
xCOMET	**0.372**	0.389	0.399	0.387
COMETDA	0.319	0.321	0.323	0.321
REG-MR-HM	0.312	0.301	0.375	0.331
REG-MR-REL	0.310	0.291	0.375	0.325
UM-MR-HM	0.338	0.332	0.391	0.354
UM-MR-REL	0.320	0.309	0.385	0.338
UM-FT-MQM	0.329	**0.448**	**0.401**	**0.393**

Table 9: Kendall's Tau correlation with human annotated MQM scores across three language pairs for QE, reference-based, and unified metrics.

Metric	en-cs	en-de	en-fi	en-ru	en-tr	en-zh	cs-en	fi-en	de-en	Average
Reference-Free (QE) Metrics										
xCOMET-NR	0.658	0.570	0.725	0.564	0.631	0.343	0.171	0.474	0.257	0.488
COMETKIWI	**0.741**	**0.593**	**0.770**	**0.617**	**0.692**	**0.378**	**0.246**	**0.515**	**0.315**	**0.541**
QE-MR	0.521	0.411	0.633	0.474	0.582	0.267	0.136	0.417	0.202	0.404
LR-LD	0.470	0.346	0.591	0.407	0.509	0.204	0.102	0.389	0.177	0.355
LR-HD	0.425	0.286	0.531	0.351	0.478	0.201	0.065	0.396	0.148	0.331
LR-LDD	0.502	0.382	0.602	0.452	0.516	0.207	0.122	0.385	0.192	0.384
LR-HDD	0.421	0.262	0.494	0.327	0.403	0.168	0.068	0.339	0.109	0.288
Reference-Based and Unified Metrics (xCOMET and Variants)										
xCOMET	0.715	**0.605**	0.744	0.604	0.694	0.401	0.222	0.509	0.308	0.533
COMETDA	**0.745**	0.598	**0.764**	**0.605**	**0.738**	**0.449**	**0.236**	**0.547**	**0.349**	**0.559**
UM-MR-HM	0.653	0.530	0.712	0.567	0.667	0.374	0.200	0.493	0.282	0.497
UM-MR-REL	0.630	0.533	0.703	0.562	0.658	0.350	0.192	0.494	0.278	0.489
REG-MR-HM	0.605	0.501	0.687	0.523	0.601	0.299	0.163	0.464	0.257	0.456
REG-MR-REL	0.612	0.518	0.666	0.504	0.614	0.309	0.151	0.438	0.236	0.450
UM-FT-MQM	0.612	0.518	0.666	0.523	0.614	0.314	0.160	0.432	0.222	0.446

Table 10: Spearman correlation with DA human annotations across selected language pairs. The average is for all 36 language pairs. While DA-trained models (e.g., COMETDA, COMETKIWI) achieve the highest scores—as expected due to fine-tuning directly on DA data. Our models (e.g., UM-MR-HM, REG-MR-HM) perform competitively despite not being trained on this supervision.



(a) 1 vs 2 (b) 2 vs 3 (c) 1 vs 3

Figure 4: AUC curves for detecting translation degradation using `QE-MR` with outputs from three MT systems in the model rotation chain: `Gemma-12B`, `Qwen-14B`, and `NLLB-distilled`. Each subplot compares model predictions between different iteration pairs—(a) 1 vs 2, (b) 2 vs 3, and (c) 1 vs 3. Higher AUC indicates better discrimination between translation quality shifts over iterations.