

# MQM Re-Annotation: A Technique for Collaborative Evaluation of Machine Translation

Parker Riley Daniel Deutsch Mara Finkelstein  
Colten Dilanni Juraj Juraska Markus Freitag  
Google

## Abstract

Human evaluation of machine translation is in an arms race with translation model quality: as our models get better, our evaluation methods need to be improved to ensure that quality gains are not lost in evaluation noise. To improve annotation quality, we experiment with a two-stage version of the current state-of-the-art translation evaluation paradigm (MQM), which we call MQM re-annotation. In this setup, an annotator reviews and edits a set of prior MQM annotations that may have come from themselves, another human annotator, or an automatic system. We demonstrate that rater behavior in re-annotation aligns with our goals, and that re-annotation results in higher-quality annotations, mostly due to finding errors that were missed during the first pass.

## 1 Introduction

A critical component of machine translation (MT) model development is evaluating model quality. Despite extensive work on automatic evaluation (Guerreiro et al., 2024; Juraska et al., 2024; Freitag et al., 2024; Kocmi and Federmann, 2023; Fernandes et al., 2023; Rei et al., 2020), human evaluation remains the gold standard for reliably determining quality. However, there are still issues affecting the quality of human evaluation, such as differences in rater behavior and task difficulty. These can negatively affect the reliability of human evaluation, and as the quality of MT systems improves, there is a risk that this evaluation noise can outweigh actual quality differences between models, leading to incorrect modeling decisions. As a result, human evaluation must also be improved to facilitate comparing ever-higher-quality systems (Freitag et al., 2021a; Kocmi et al., 2024a, 2023).

The current state-of-the-art human evaluation framework for MT is Multidimensional Quality Metrics (MQM), originally proposed by Lommel et al. (2014). Freitag et al. (2021a) adapted MQM

for MT, arguing that the use of fine-grained annotation by expert raters improved the quality of human evaluation. In MQM, expert raters identify error spans within a translation, assigning a severity and hierarchical error category to each. Follow-up work has examined other facets of evaluation that can be used to enhance quality, including comparative judgment (Song et al., 2025), careful rater assignment (Riley et al., 2024), and rater calibration (Kocmi et al., 2022).

In this work we explore both human-machine collaboration and human-human collaboration, with the goal of increasing quality of MQM annotation for machine translation evaluation. We do this using *re-annotation*, where a completed MQM annotation of a document is given to an expert human rater who can delete or modify any previously-marked error spans in addition to adding new ones. The prior rater can either be the same human rater, a different human rater, or an automatic system.

We release our data to the public to encourage further analysis.<sup>1</sup>

## 2 Research Questions and Findings

In this section, we outline the research questions we seek to answer in this work and summarize the findings. Evidence supporting these findings is presented in Section 5, with relevant context and background covered in the intervening sections.

### 2.1 How much do raters change prior annotations?

Our first goal is to understand what happens when raters perform re-annotation: how often do they change/delete prior errors or add new ones? Further, to what extent does this depend on whether those prior errors came from themselves, a different human, or an automatic system?

<sup>1</sup><https://github.com/google/wmt-mqm-human-evaluation/tree/main/generalMT2023/>

We find that rater behavior is highly sensitive to the source of the prior errors, and trends align with our expectations: raters perform the fewest overall modifications to their own annotations, more to other humans' annotations, and even more modifications to automatic annotations. We also find evidence suggesting that a primary benefit of re-annotation is that it allows raters the opportunity to find errors that were missed initially.

## 2.2 Do raters detect artificial error spans?

One potential risk with re-annotation is that raters may overly trust the prior errors and focus solely on adding new ones. To examine this, we injected artificial error spans and measured how often raters modified or deleted them.

We found that, on average, raters removed most of the artificial errors, but that a minority of raters accepted these errors as-is with high frequency.

## 2.3 Does re-annotating human annotations improve their quality?

The prior two research questions establish that re-annotation is *reasonable*, in that re-annotators behave in ways that are consistent with our expectations, but we also wish to understand whether it is *beneficial*, in that it improves rating quality. Because defining a ground truth is difficult in human evaluation research, we measure whether re-annotation can allow two disjoint groups of human raters to achieve higher cross-group agreement.

We find that re-annotating human MQM ratings (either by the same rater or a different human rater) increases their quality.

## 2.4 Do automatic prior annotations improve human annotation quality?

As a cheaper alternative to human-human re-annotation, we examine whether re-annotating *automatic* ratings is beneficial over standard MQM annotation from scratch. Based on the previous research question, we can use our re-annotated data as the proxy ground truth to measure improvement.

We find that providing raters with prior annotations from LLM-based automatic systems improves rating quality over from-scratch annotation, at no additional human annotation cost.

## 2.5 Which kind of re-annotation is best?

This is a natural followup to the two prior research questions. We use the same technique as in the two

Source (en)	Translation (de)	Evaluations
Kazakh rescuers pull out 7 survivors and bodies of 19 deceased in Türkiye: 11 February 2023, 12:15	Kasachstanische Rettungskräfte holen 7 Überlebende und 19 Leichen aus der Türkei heraus: 11. Februar 2023, 12:15	<p>Click on the start of an error span not yet marked</p> <p>Major severity [M] Minor severity [m]</p> <p>Accuracy: Mistranslation: <b>holen</b></p> <p>Accuracy: Mistranslation: <b>aus der Türkei heraus</b></p>
<p>The Kazakhstani rescue team pulled out seven quake survivors and the bodies of 19 people, including one kid, in Türkiye. First medical aid is also provided on a round-the-clock basis.</p> <p>Die Rettungskräfte in Kasachstan haben sieben Überlebende und die Leichen von 19 Personen, darunter ein Kind. Erste medizinische Hilfe wird rund um die Uhr geleistet.</p>		

Figure 1: Anthea re-annotation interface.

prior research questions (human re-annotation data as the proxy ground truth).

We find that human-human re-annotations are generally higher quality than re-annotated automatic annotations when two different human raters perform the annotation, though the approximately doubled human labor cost means that the choice of which to use is situation-dependent.

## 3 MQM and Re-Annotation

In the Multidimensional Quality Metrics (MQM) framework for human evaluation of machine translation (MT), expert annotators identify error spans within translations and assign a hierarchical error category (e.g. Fluency/Grammar or Accuracy/Mistranslation) and severity (Major or Minor<sup>2</sup>). Instead of raters providing scalar scores, scores are calculated by applying a weighting scheme based primarily on severity to each error and summing over the segment; see Freitag et al. (2021a) for details. All source and target segments for a single system's translation of a single document are presented to the rater on the same screen.

MQM re-annotation functions the same as standard MQM annotation, except that some error spans are already present when raters first load the interface. Raters have the ability to modify any component of a prior error (span start/end, category, severity), delete it entirely, and add new errors.

We use the open-source Anthea<sup>3</sup> framework to collect MQM ratings, including re-annotations. The rating interface is shown in Figure 1.

<sup>2</sup>Major error significantly degrade the quality of the translation, while Minor errors do not.

<sup>3</sup><https://github.com/google-research/google-research/tree/master/anthea>

## 4 Experimental Setup

### 4.1 Data

We use translation data from the WMT 2023 Metrics Shared Task (Freitag et al., 2023) in English-German and Chinese-English. For budget reasons, we only use data from the News domain. Information about the dataset size is shown in Table 1.

The data contains 3 initial human MQM annotations per segment, and two automatic annotations from different systems (Section 4.3). Each of these annotations was then re-annotated by one of the same 3 human raters who initially rated that segment. Further details are in Section 4.2.

The human annotators are professional translators who regularly perform MQM annotation. They were paid fair market wages for their work.

	Chinese-English	English-German
# Source Segments	247	100
# Source Documents	25	29
# Systems	16	13
# Raters	8	10

Table 1: Counts of segments, documents, and systems in each language pair. One system was a human reference. “# Raters” refers to the total number of unique raters; see Section 4.2 for a description of how many ratings of which types were collected.

### 4.2 Re-Annotation Settings

We use  $h_i$ ,  $i \in [1, N]$ , to refer to MQM annotations from our  $N$  human raters for a particular language pair, with no re-annotation. We use  $a_x$ ,  $x \in \{1, 2\}$  to refer to MQM annotations produced by an automatic system. For re-annotation, we adopt notation from conditional probability: for example, we use  $h_i|h_j$  to refer to the MQM annotations from human rater  $i$  re-annotating those from human rater  $j$ .

We define the following (re-)annotation types:

- **Single:** Initial annotations by a single human rater<sup>4</sup>:  $h_i$ ,  $i \in [1, N]$
- **Self:** Re-annotation done by the same human rater who did the initial annotation:  $h_i|h_i$
- **Other:** Re-annotation done by a human rater who is different from the initial rater:  $h_j|h_i$ ,  $i \neq j$

<sup>4</sup>We do not consider automatic annotations in isolation in this work.

- **Auto:** Re-annotation done by a human rater when the initial annotations came from an automatic system:  $h_i|a_x$

	Chinese-English	English-German
Single	11856	3900
Self	3952	1300
Other	7904	2600
Auto	7410	2400

Table 2: Number of segment annotations for each (re-)annotation setting and language pair.

Due to budget constraints, we were unable to collect data from all possible combinations of raters for every source document. Each document was rated by 3 human raters and 2 automatic systems, and these initial annotations were distributed among the same 3 human raters as follows: one rater was randomly selected to re-annotate their own ratings (the Self setting), and the other two re-annotated each other’s ratings (the Other setting). Independently, 2 of the 3 raters were randomly selected to each re-annotate a different automatic system. All random assignments were made separately for each document, and all system outputs of all segments for each document shared its assignment. However, automatic annotations were not available for the human reference system in the data (“refA”), so the Auto setting is somewhat smaller than the Other setting. When re-annotating, raters were not told the origin of the initial annotations. Table 2 details how many segment annotations our data contains for each setting. These counts reflect the product of the number of segments, the number of systems (see Table 1), and the number of annotations in each setting: 3 for Single, 1 for Self, and 2 for Other and Auto (with the human reference “system” excluded from the Auto setting).

### 4.3 Automatic Systems

We experiment with two automatic MQM annotation systems: GEMBA-MQM (Kocmi and Federmann, 2023), which is based on prompting GPT-4 (OpenAI, 2024), and a variant of AutoMQM (Fernandes et al., 2023) based on fine-tuning Gemini 1.0 Pro (Gemini Team, 2024). The AutoMQM model was fine-tuned on MQM data from the WMT Metrics Shared Task, years 2020 through 2022 (Mathur et al., 2020; Freitag et al., 2021b, 2022). Both models are reference-free, taking as input only the source segment and translation hypothesis.

## 4.4 Artificial Spans

To test for over-reliance on prior annotations, we selected one document in each language pair to augment with artificial error spans. For each system translation of this document, and for each segment, we randomly selected a one- or two-token span that did not overlap with any errors in the three human initial annotations and assigned a Major severity and a random category. The Major severity was chosen to increase our confidence that the artificial errors were indeed incorrect. These artificial error spans were marked in addition to the actual spans identified by the rater who marked the most total errors over the chosen document’s translations. This was done to maximize the number of legitimate error spans in order to make the artificial spans less obvious. The chosen document was only used to analyze rater behavior on the artificial spans, and was not included in any other analyses in this work.

## 4.5 Agreement Metrics

### 4.5.1 Score Agreement

To measure inter-annotator agreement of MQM scores, we use Pairwise Ranking Agreement (PRA) (Song et al., 2025; Deutsch et al., 2023). PRA essentially measures how often two sets of annotations agree on the relative ranking of two translations, treated as a 3-way classification task: A is either better than, worse than, or equal to B. This is summarized in Equation 1:

$$\text{PRA}_X(\alpha, \beta) = \frac{1}{\binom{|X|}{2}} \sum_{x_1, x_2 \in X} \mathbb{1}[\alpha(x_1, x_2) = \beta(x_1, x_2)] \quad (1)$$

where  $X$  is a set of translation system scores (possibly averages over multiple segments),  $\alpha$  and  $\beta$  are ranking functions that indicate which input is better (or whether they are tied) according to a given set of human annotations,  $\binom{|X|}{2}$  is the number of pairs of system scores in  $X$ , and  $\mathbb{1}$  is the indicator function. We primarily use the Group-by-Item segment-level setup of Deutsch et al. (2023), where PRA is calculated over all system translations of a given source segment, and the average of this value over all source segments is reported; we refer to this as “segment-level pairwise agreement”.

We set up our data collection and correlation analysis in such a way that two sets of annotations from the same human rater are never directly

compared when calculating agreement, as doing so would risk biasing the results based on rater-specific behaviors. We designed the setup to primarily focus on segment-level agreement, where the agreement on each segment is calculated between disjoint sets of raters. Segment-level agreement has been shown to be better at discriminating metrics than system-level agreement (Deutsch et al., 2023; Freitag et al., 2022). Despite this, it is still common to report system-level agreement, but our setup prevents us from doing this because all annotators participated in all annotation settings. As an approximation, we instead calculate “rater-level” agreement, by partitioning our data based on which rater was selected to re-annotate their own ratings, and reporting the agreement of two settings as the average agreement over each part.

### 4.5.2 Span Agreement

To measure inter-annotator agreement of error spans and severity markings, we use character-level F1 scores (abbreviated “character F1”), introduced by Blain et al. (2023). Character F1 involves comparing whether two annotations agree on whether a character is or is not included in an error span, with half-points awarded when both annotations include the character in a span but disagree on the severity.

## 5 Results

### 5.1 How much do raters change prior annotations?

Table 3 reports the relative frequencies of deleting/changing/keeping/adding errors during re-annotation. In both language pairs, non-keep rates increase when moving from Self to Other to Auto. These results validate our hypothesis that any given rater should generally agree with themselves more than a different human, and with a different human more than with an automatic rater. The Added % for the Auto setting is strikingly higher than in the other two settings, leading us to compare the average total number of errors identified over the entire dataset by our human and automatic raters. We found that human annotators identified 2.7 times as many errors during initial annotation as automatic systems did, in both language pairs.

Additionally, we note that re-annotation increases the average number of annotated error spans in all settings. This is true even in the Self setting, where one might assume that raters would change very little. This provides evidence that a pri-

	Chinese-English			English-German		
	Self	Other	Auto	Self	Other	Auto
Deleted %	10.9	25.3	28.7	10.2	20.9	24.9
Changed %	7.8	12.8	23.3	6.4	11.7	27.6
Kept %	81.3	61.9	48.0	83.4	67.3	47.5
Added %	30.1	59.2	173.8	41.0	54.6	202.1

Table 3: Number of deleted/changed/kept/added errors, expressed as percentages of the total number of prior errors, by language pair and re-annotation setting. Deleted %, Changed %, and Kept % add to 100% within a column. Added % can be higher than 100% if the number of newly added errors exceeds the number of prior errors. Percentages are macro-averages of individual re-annotator performances.

	Chinese-English				English-German			
	Self	Other	Auto	Single	Self	Other	Auto	Single
Minor %	86.0	78.7	65.8	73.2	75.6	73.1	69.9	68.1

Table 4: Percentages of errors added during re-annotation (or in the Single setting, initial annotation) that had Minor severity.

	Chinese-English			English-German		
	Self	Other	Auto	Self	Other	Auto
Minor Kept %	77.8	60.0	40.2	81.3	66.1	45.7
Major Kept %	92.7	68.9	52.2	87.0	68.9	52.6

Table 5: Percentage of prior errors that were left unchanged by re-annotators, separated by severity of the prior error.

mary benefit of re-annotation is **finding missed errors** (as opposed to removing/changing incorrectly-annotated errors).

Motivated by this finding, we examined the severity levels of errors added in re-annotation, compared to in the original annotations. The results are shown in Table 4, which shows that, in both language pairs, human re-annotation (Self, Other) leads to a higher proportion of Minor errors than in the original annotations (Single). This is consistent with the hypothesis that Major errors are more salient and thus less likely to be missed, though we note that plenty of Major errors are still found in re-annotation. However, in the Auto setting, the proportion of Minor errors is noticeably lower than the human re-annotation settings, and for Chinese-English it is even lower than in the Single setting. This may indicate that there is a mismatch between the relative salience of errors between human and automatic annotators.

To further investigate the relationship between error severity and salience, we investigated whether an initial error’s severity affects its likelihood of being modified during re-annotation. These results are shown in Table 5. In both language pairs and all settings, Major errors were more likely to be left unchanged. This suggests that not only are Major errors more likely to be identified on the first pass

(at least with human initial annotators), but they are also more likely to be agreed with by re-annotators.

Although finding missed errors appears to be the primary benefit of re-annotation, we also examined errors that were changed in re-annotation to see whether raters primarily changed the severity, category, or span. Results are shown in Table 6. We find a different trend between our Chinese-English and English-German raters, where in Chinese-English re-annotators in the Self setting mostly changed severity, while in the Other setting mostly changed category. For English-German, the most common edit was to category in both the Self and Other settings. While we do not have a good explanation for the elevated number of severity changes in the Self setting for Chinese-English, the rest of the human re-annotation results could point to lack of clarity in category descriptions in the MQM instructions. Looking next at the Auto setting, both language pairs show that most changes are to error span boundaries, indicating that automatic annotation systems struggle with precisely demarcating errors. Manual analysis reveals that automatically-identified error spans are often significantly longer than human-identified ones.

	Chinese-English			English-German		
	Self	Other	Auto	Self	Other	Auto
Severity Changed %	52.2	38.2	17.8	32.1	39.3	30.9
Category Changed %	35.8	56.8	21.2	63.5	68.7	37.0
Span Changed %	33.1	34.6	81.4	24.2	20.8	67.3

Table 6: Percentages of changed errors where the severity, category, or span was edited (because multiple edits can be made to each error, percentages will sum to more than 100%), by re-annotation setting.

	Chinese-English	English-German
Deleted %	82.3	70.8
Changed %	4.9	4.4
Kept %	12.8	24.8

Table 7: Percentage of artificial error annotations that were deleted/changed/kept by re-annotators. Percentages are macro-averaged over individual re-annotators.

## 5.2 Do raters detect artificial error spans?

Table 7 shows the rates at which re-annotators deleted, changed, or kept artificial error spans. While re-annotators usually did not keep artificial error spans unchanged, they did so more than we expected. We expected the percentage of kept errors to be nearly 0 because we selected artificial error spans that did not overlap with any prior human annotations. When examining per-rater performance, we found that a few outlier raters were keeping prior errors at concerningly high rates (78.8% for one English-German rater). The median percentage of kept artificial errors is 8.8% for Chinese-English and 13.5% for English-German.

## 5.3 Does re-annotating human annotations improve their quality?

Tables 8 and 9 report the agreement between different (re-)annotation settings using character F1 and average segment-level PRA, respectively (see Section 4.5 for details on agreement metrics). Under both metrics, the agreement between two annotations in the Single setting (e.g.  $h_i$  vs.  $h_j$ ) is increased by replacing either side with a re-annotation (e.g.  $h_i|h_i$  vs.  $h_j$ ), and increased further by replacing both sides (e.g.  $h_i|h_i$  vs.  $h_j|h_k$ ).

While it is unclear how to calculate a  $p$ -value for comparing two character F1 scores, for pairwise agreement we use a paired permutation test with 10,000 permutations to compare pairs of values in Table 9. We find that in both language pairs, the increase in pairwise agreement from Single vs. Single to Self vs. Other is statistically significant with a  $p$ -value of less than 0.001.

When comparing “rater-level” agreement (see Section 4.5.1) in Table 10, the overall trend is the same except for the left two columns, where agreement with  $h_j$  is higher than with  $h_j|h_k$ . Despite this, the ultimate conclusion that Single vs. Single agreement is lower than Self vs. Other still holds.

Because all agreement scores are calculated between annotations stemming from two disjoint sets of human annotators (avoiding bias), we conclude that the increased agreement is an indication that re-annotation improves human rating quality.

## 5.4 Do automatic prior annotations improve human annotation quality?

Having established that re-annotating human ratings improves their quality, we can use our re-annotations as proxies for the unknown “ground truth” MQM annotations and investigate whether providing raters with automatic prior annotations improves their rating quality. Tables 11 and 12 show average segment-level agreement between (re-)annotation settings, with a goal of determining whether the Auto setting ( $h_i|a_x$ ,  $h_j|a_y$ ) improves over the Single setting ( $h_i$ ,  $h_j$ ) when using either the Self ( $h_i|h_i$ ) or Other ( $h_j|h_k$ ) setting as the ground truth. Note that, in order to ensure that all comparisons within a language pair use the exact same data, these tables use 2/3 of the available data, discarding the third where neither of the initial ratings from the automatic systems were distributed to the rater who re-annotated their own ratings ( $h_i$ ). In both language pairs, under both metrics, and for both  $h_i|a_x$  and  $h_j|a_y$ , we see that providing humans with automatic prior annotations improves their agreement with human-only re-annotation, indicating an increase in quality; the only exception is in Table 8 for Chinese-English where  $h_j|a_y$  ties  $h_j$  in terms of span-aware agreement with  $h_i|h_i$ .

For the pairwise agreement scores in Table 12, we calculate statistical significance as in Section 5.3, and the results mirror the conclusions based on character F1 above: in both language pairs, for both  $h_i|a_x$  and  $h_j|a_y$ , the increase in

	Chinese-English				English-German			
	$h_j$ (Single)	$h_j h_k$ (Other)	$h_k$ (Single)	$h_k h_j$ (Other)	$h_j$ (Single)	$h_j h_k$ (Other)	$h_k$ (Single)	$h_k h_j$ (Other)
$h_i$ (Single)	0.333	0.356	0.293	0.343	0.329	0.360	0.343	0.354
$h_i h_i$ (Self)	0.351	0.368	0.309	0.363	0.352	0.391	0.359	0.387

Table 8: Character F1 scores between different human-only annotation settings.

	Chinese-English				English-German			
	$h_j$ (Single)	$h_j h_k$ (Other)	$h_k$ (Single)	$h_k h_j$ (Other)	$h_j$ (Single)	$h_j h_k$ (Other)	$h_k$ (Single)	$h_k h_j$ (Other)
$h_i$ (Single)	0.534	0.557	0.527	0.543	0.582	0.611	0.567	0.614
$h_i h_i$ (Self)	0.544	0.575	0.539	0.564	0.599	0.658	0.603	0.647

Table 9: Average segment-level pairwise agreement scores between different human-only annotation settings.

agreement with human-only re-annotation gained by providing humans with automatic prior annotations is statistically significant ( $p < 0.05$ ), except in Chinese-English where  $h_j|a_y$  and  $h_j$  are not significantly different in agreement with  $h_i|h_i$ .

The rater-level analysis in Table 13 shows mostly similar trends as Table 12, except for Chinese-English when moving from left to right in each row, where agreement with  $h_j$  is unusually high. This mirrors the findings in Table 9, indicating unusually high agreement with  $h_j$  that does not seem to be shared with  $h_k$ . It is possible that this is a result of an interaction between idiosyncratic rater behaviors, which the Other setting is most suited to addressing (see Section 5.1); we therefore focus on using  $h_j|h_k$  as the “ground truth”, where the conclusion is the same at the segment and rater level: automatic prior annotations increase agreement over the Single setting.

Critically, these results illustrate that automatic prior annotations can increase human rating quality without any additional human annotation cost.<sup>5</sup>

See Appendix A.1 for additional comparisons with the initial automatic annotations.

### 5.5 Which kind of re-annotation is best?

We re-use Table 11, Table 12, and Table 13 to answer this question. First, to compare the Auto and Other settings, we use Self as the ground truth and see that for both language pairs, both agreement metrics, and both analysis levels for PRA, the Other setting achieves higher agreement, indicating that a collaboration of two humans is stronger than a human-machine collaboration. Next, to compare

the Auto and Self settings, we use Other as the ground truth and find more mixed results. Using character F1, Auto outperforms Self in both language pairs. Using PRA, Self outperforms Auto at both analysis levels in English-German, but in Chinese-English they are effectively tied at the segment level, and Auto has a slight edge at the rater level.

We therefore conclude that while the Auto and Self settings may be comparable, the Other setting has a distinct advantage. As discussed in Section 5.4, we hypothesize that the interaction of two human raters minimizes the impact of idiosyncratic rater behavior, resulting in the highest quality MQM annotations. However, there is a tradeoff where that increased quality comes at approximately doubled human labor cost, making it best suited for scenarios where individual ratings need to be of utmost quality, such as for metric training and evaluation.

## 6 Related Work

While our work focuses on improving evaluation quality, a notable recent work focusing instead on evaluation cost is Error Span Annotation (ESA) (Kocmi et al., 2024b). ESA is highly similar to MQM, but the primary cost-saving difference is the removal of error category information. This reduces annotation time, and Kocmi et al. (2024b) claim that it also permits the use of non-expert raters.

The most closely related work to ours is that of Zouhar et al. (2025), who essentially do re-annotation of automatic prior annotations in the ESA framework. The largest difference between their work and ours is that we also investigate re-annotation of *human* prior annotations, which we

<sup>5</sup>We found that re-annotation of automatic systems took 12% less time than initial annotation for Chinese-English and 2% less time for English-German, on average.

	Chinese-English				English-German			
	$h_j$ (Single)	$h_j h_k$ (Other)	$h_k$ (Single)	$h_k h_j$ (Other)	$h_j$ (Single)	$h_j h_k$ (Other)	$h_k$ (Single)	$h_k h_j$ (Other)
$h_i$ (Single)	0.815	0.811	0.790	0.824	0.794	0.823	0.790	0.826
$h_i h_i$ (Self)	0.853	0.828	0.808	0.851	0.824	0.860	0.819	0.849

Table 10: Average rater-level pairwise agreement scores between different human-only annotation settings.

	Chinese-English			English-German		
	$h_j$ (Single)	$h_j a_y$ (Auto)	$h_j h_k$ (Other)	$h_j$ (Single)	$h_j a_y$ (Auto)	$h_j h_k$ (Other)
$h_i$ (Single)	0.315	0.330	0.332	0.353	0.383	0.368
$h_i a_x$ (Auto)	0.352	0.380	0.375	0.411	0.422	0.428
$h_i h_i$ (Self)	0.337	0.337	0.348	0.384	0.401	0.412

Table 11: Character F1 scores between different annotation settings.

	Chinese-English			English-German		
	$h_j$ (Single)	$h_j a_y$ (Auto)	$h_j h_k$ (Other)	$h_j$ (Single)	$h_j a_y$ (Auto)	$h_j h_k$ (Other)
$h_i$ (Single)	0.511	0.510	0.530	0.574	0.598	0.579
$h_i a_x$ (Auto)	0.530	0.546	0.552	0.619	0.631	0.630
$h_i h_i$ (Self)	0.528	0.530	0.553	0.598	0.632	0.641

Table 12: Average segment-level pairwise agreement scores between different (re-)annotations.

	Chinese-English			English-German		
	$h_j$ (Single)	$h_j a_y$ (Auto)	$h_j h_k$ (Other)	$h_j$ (Single)	$h_j a_y$ (Auto)	$h_j h_k$ (Other)
$h_i$ (Single)	0.766	0.752	0.766	0.730	0.773	0.776
$h_i a_x$ (Auto)	0.833	0.799	0.814	0.771	0.815	0.805
$h_i h_i$ (Self)	0.829	0.792	0.804	0.777	0.806	0.852

Table 13: Average rater-level pairwise agreement scores between different (re-)annotations.

show can be used to create higher-quality reference annotations. It also provides a novel, direct way to examine inter-annotator agreement by directly comparing re-annotators’ behavior on their own ratings to their behavior on others’. Additionally, we expect that our use of expert annotators reduces the available headroom for improving rating quality, making the observed increases from re-annotation (both of human and automatic prior annotations) more noteworthy.

Other works have attempted to increase the quality of machine translation evaluation, including MQM. Song et al. (2025) leveraged side-by-side MQM annotation (where two translations of the same source were shown on-screen and otherwise annotated normally) to improve rating consistency. Riley et al. (2024) demonstrated that careful rater assignment can mitigate noise from differences in rater behavior. Knowles (2021) showed that score normalization, a technique intended to address dif-

ferences in rater behavior, can be harmful if applied in the wrong setting. Saldías Fuentes et al. (2022) proposed a technique for selecting test data to send for human evaluation that improved reliability in a budget-constrained setting.

Re-annotation is conceptually similar to translation post-editing, where the translation itself is updated in a two-stage process as opposed to the *annotation* of the translation as in this work. In this way, our work is loosely reminiscent of Liu et al. (2024), who investigated human-only and human-machine collaboration for translation post-editing.

## 7 Conclusion

In this work, we presented an analysis of MQM re-annotation, exploring settings where the initial annotator was the re-annotator themselves, a different human, or an automatic system. We validated that the re-annotation task elicits reasonable rater behavior, and provided evidence that it leads to higher-

quality annotations. In the case of re-annotation of automatic ratings, this quality improvement comes at no additional human annotation cost. As for re-annotation of human ratings, the improved quality makes it a candidate for creating test sets for meta-evaluation of automatic metrics, especially span-level metrics where the quality of the ground-truth spans is critically important.

## Limitations

All annotation and re-annotation in this work was exclusively performed by professional translators who regularly do MQM annotation, meaning that our findings may not apply to less-experienced annotators. Additionally, we only experimented with two automatic MQM systems that both have reasonable quality; it is possible that re-annotating a lower-quality system would not outperform from-scratch human-only annotation. Finally, our data is exclusively from the News domain, leaving the possibility that re-annotators might perform differently on different kinds of text.

## References

- Frederic Blain, Chrysoula Zerva, Ricardo Rei, Nuno M. Guerreiro, Diptesh Kanojia, José G. C. de Souza, Beatriz Silva, Tânia Vaz, Yan Jingxuan, Fatemeh Azadi, Constantin Orasan, and André Martins. 2023. [Findings of the WMT 2023 shared task on quality estimation](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 629–653, Singapore. 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.
- Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. 2023. [The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 1066–1083, Singapore. Association for Computational Linguistics.
- 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, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. [Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics.
- Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. [Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021b. [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.
- Gemini Team. 2024. [Gemini: A family of highly capable multimodal models](#). *Preprint*, arXiv:2312.11805.
- 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.
- 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.
- Rebecca Knowles. 2021. [On the stability of system rankings at WMT](#). In *Proceedings of the Sixth Conference on Machine Translation*, pages 464–477, Online. 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, and 3 others. 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, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, and 2 others. 2023. [Findings of the 2023 conference on machine translation \(WMT23\): LLMs are here but not quite there yet](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.
- Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. [Findings of the 2022 conference on machine translation \(WMT22\)](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Tom Kocmi and Christian Federmann. 2023. [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, 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.
- Zhongtao Liu, Parker Riley, Daniel Deutsch, Alison Lui, Mengmeng Niu, Apurva Shah, and Markus Freitag. 2024. [Beyond human-only: Evaluating human-machine collaboration for collecting high-quality translation data](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 1095–1106, Miami, Florida, USA. Association for Computational Linguistics.
- Arle Lommel, Aljoscha Burchardt, and Hans Uszkor-eit. 2014. [Multidimensional quality metrics \(mqm\): A framework for declaring and describing translation quality metrics](#). *Tradumàtica: tecnologies de la traducció*, 0:455–463.
- Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020. [Results of the WMT20 metrics shared task](#). In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online. Association for Computational Linguistics.
- OpenAI. 2024. [Gpt-4 technical report](#). *Preprint*, arXiv:2303.08774.
- 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.
- Parker Riley, Daniel Deutsch, George Foster, Viresh Ratnakar, Ali Dabirmoghaddam, and Markus Freitag. 2024. [Finding replicable human evaluations via stable ranking probability](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4908–4919, Mexico City, Mexico. Association for Computational Linguistics.
- Belén Saldías Fuentes, George Foster, Markus Freitag, and Qijun Tan. 2022. [Toward more effective human evaluation for machine translation](#). In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 76–89, Dublin, Ireland. 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.
- Vilém Zouhar, Tom Kocmi, and Mrinmaya Sachan. 2025. [AI-assisted human evaluation of machine translation](#). *Preprint*, arXiv:2406.12419.

## A Appendix

### A.1 Initial Autorater Performance

Here we present expanded versions of Tables 4, 11, 12, and 13 that include results from the initial automatic labeler models themselves; in these tables we refer to this setting as LLM as both models we consider are LLM-based. These additional results are shown in Tables 14, 15, 16, and 17.

Table 14 shows that the automatic systems produce a much smaller proportion of Minor errors in Chinese-English than in English-German, where the proportion is on par with the human-based annotation settings.

Tables 15, 16, and 17 show that the automatic systems have high agreement with each other in Chinese-English, but in English-German their agreement with each other is comparable to other agreement scores.

	Chinese-English					English-German				
	Self	Other	Auto	Single	LLM	Self	Other	Auto	Single	LLM
Minor %	86.0	78.7	65.8	73.2	34.8	75.6	73.1	69.9	68.1	73.2

Table 14: Percentages of errors added during re-annotation (or in the Single and LLM settings, initial annotation) that had Minor severity. The LLM setting represents the errors added by the automatic labeler models.

	Chinese-English				English-German			
	$a_y$ (LLM)	$h_j$ (Single)	$h_j a_y$ (Auto)	$h_j h_k$ (Other)	$a_y$ (LLM)	$h_j$ (Single)	$h_j a_y$ (Auto)	$h_j h_k$ (Other)
$a_x$ (LLM)	0.373	0.286	0.339	0.280	0.250	0.269	0.288	0.287
$h_i$ (Single)	0.278	0.315	0.330	0.332	0.255	0.353	0.383	0.368
$h_i a_x$ (Auto)	0.324	0.352	0.380	0.375	0.281	0.411	0.422	0.428
$h_i h_i$ (Self)	0.283	0.337	0.337	0.348	0.271	0.384	0.401	0.412

Table 15: Character F1 scores between different annotation settings, with initial automatic labeler performance (LLM).

	Chinese-English				English-German			
	$a_y$ (LLM)	$h_j$ (Single)	$h_j a_y$ (Auto)	$h_j h_k$ (Other)	$a_y$ (LLM)	$h_j$ (Single)	$h_j a_y$ (Auto)	$h_j h_k$ (Other)
$a_x$ (LLM)	0.602	0.484	0.519	0.470	0.559	0.548	0.570	0.518
$h_i$ (Single)	0.469	0.511	0.510	0.530	0.519	0.574	0.598	0.579
$h_i a_x$ (Auto)	0.508	0.530	0.546	0.552	0.514	0.619	0.631	0.630
$h_i h_i$ (Self)	0.466	0.528	0.530	0.553	0.525	0.598	0.632	0.641

Table 16: Average segment-level pairwise agreement scores between different (re-)annotations, with initial automatic labeler performance (LLM).

	Chinese-English				English-German			
	$a_y$ (LLM)	$h_j$ (Single)	$h_j a_y$ (Auto)	$h_j h_k$ (Other)	$a_y$ (LLM)	$h_j$ (Single)	$h_j a_y$ (Auto)	$h_j h_k$ (Other)
$a_x$ (LLM)	0.801	0.789	0.754	0.762	0.729	0.714	0.745	0.717
$h_i$ (Single)	0.733	0.766	0.752	0.766	0.767	0.730	0.773	0.776
$h_i a_x$ (Auto)	0.792	0.833	0.799	0.814	0.764	0.771	0.815	0.805
$h_i h_i$ (Self)	0.778	0.829	0.792	0.804	0.768	0.777	0.806	0.852

Table 17: Average rater-level pairwise agreement scores between different (re-)annotations, with initial automatic labeler performance (LLM).