Enhancing Human Evaluation in Machine Translation with Comparative Judgment

Yixiao Song[♠*] **Parker Riley**[◇] **Daniel Deutsch**[◇] **Markus Freitag**[◇]
[♠]Manning College of Information and Computer Sciences, UMass Amherst
[◇]Google
yixiaosong@umass.edu {prkriley, dandeutsch, freitag}@google.com

Abstract

Human evaluation is crucial for assessing rapidly evolving language models but is influenced by annotator proficiency and task design. This study explores the integration of comparative judgment into human annotation for machine translation (MT) and evaluates three annotation setups—point-wise Multidimensional Quality Metrics (MQM), side-by-side (S×S) MQM, and its simplified version S×S relative ranking (RR). In MQM, annotators mark error spans with categories and severity levels. S×S MQM extends MQM to pairwise error annotation for two translations of the same input, while S×S RR focuses on selecting the better output without labeling errors.

Key findings are: (1) the S×S settings achieve higher inter-annotator agreement than MQM; (2) S×S MQM enhances inter-translation error marking consistency compared to MQM by, on average, 38.5% for explicitly compared MT systems and 19.5% for others; (3) all annotation settings return stable system rankings, with S×S RR offering a more efficient alternative to (S×S) MQM; (4) the S×S settings highlight subtle errors overlooked in MQM without altering absolute system evaluations.

To spur further research, we release the triply annotated datasets comprising 377 ZhEn and 104 EnDe annotation examples, each covering 10 systems.[1]

1 Introduction

With the rapid improvement of large language models' capabilities, automatic evaluation metrics have struggled to reliably measure their quality (Karpinska and Iyyer, 2023; Pham et al., 2024). As a result, human evaluation continues to play a vital role in assessing models' performance.

Human annotations can be influenced by several difficult-to-control factors, such as annotators' proficiency and their relative leniency or stringency (Lu et al., 2025). Annotator proficiency can be managed by hiring experts (Karpinska et al., 2021; Krishna et al., 2023). Varying degrees of leniency or stringency can be mitigated by carefully assigning tasks to annotators in a structured manner (Riley et al., 2024). However, other factors can affect rater behavior, such as the specific annotation task used to measure quality (Belz and Kow, 2010).

This work investigates the influence of annotation settings on annotator behavior and results by using Chinese to English (ZhEn) and English to German (EnDe) machine translation (MT) as a case study. It examines three annotation settings: (1) the state-of-the-art point-wise MT annotation setup **MQM** (Lommel et al., 2014b; Freitag et al., 2021a) where annotators see one translation at a time and identify errors with category and severity assignment of each, (2) **side-by-side** (S×S) **MQM**, where annotators see two translations of the same input at a time and give fine-grained error annotations as MQM, (3) **S×S relative ranking** (RR), where annotators see two translations and decide which one is better, without error annotation. The latter two settings incorporate comparative judgment (Thustone, 1927), a pair-wise setting that allows annotators to make relative assessments between system outputs; in S×S RR, it assists annotators in making comparisons, while in S×S MQM it helps them detect errors more easily, particularly those appearing in only one output. Comparative judgment has been shown to reduce subjectivity and enhance consistency in quality judgments (Karpinska et al., 2021; de Moira et al., 2022; Jones and Davies, 2024). The three settings are illustrated in Figure 1.

This work meta-evaluates human annotation results from the studied annotation settings across five aspects: inter-annotator agreement, inter-

*Work done during an internship at Google.

[1]Data will be available at https://github.com/google/wmt-mqm-human-evaluation/tree/main/generalMT2023. In EnDe, one system is annotated twice (see Section 3.2).

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 20536–20551
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

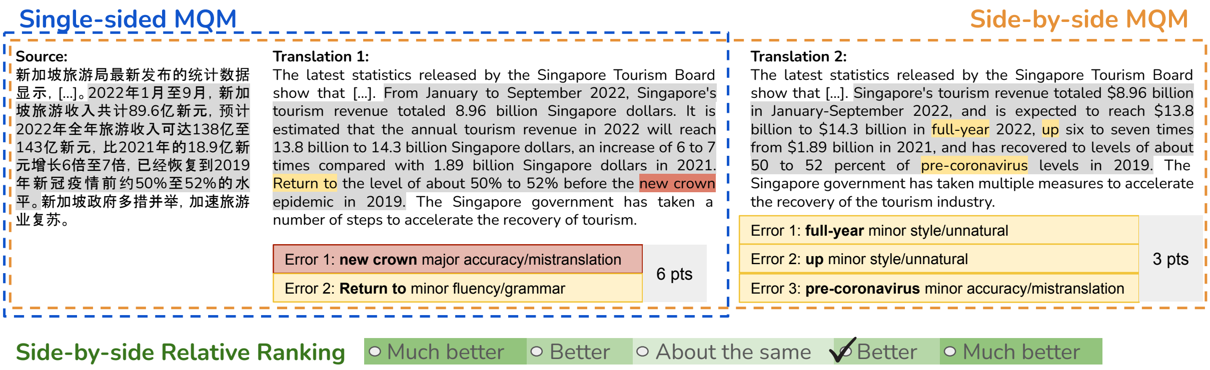


Figure 1: Illustration of the three annotation settings studied in this work (Section 2.4). The grey-highlighted text is the segment to be annotated within their context. In **single-sided** and **side-by-side** MQM, annotators mark error spans and assign error category with severity. The score of a segment/document is determined by the category and severity of its error(s). In **side-by-side relative ranking**, annotators read two translations and choose the (much) better side or decide if they tie, without labelling errors. The scoring scheme of each setting is in Section 2.5.

translation error marking consistency, segment- and system-level quality rankings, and error distribution in MQM and S×S MQM. Overall, our human annotation results reveal the following relative strength/weakness of each protocol. **Comparative judgment improves inter-annotator agreement**, especially in S×S MQM, compared to point-wise MQM. **Comparative judgment boosts inter-translation error marking consistency**, with average increases of 38.05% for systems that were shown side-by-side, and 19.5% for system pairs that were not. For MT system ranking, **S×S MQM proved more reliable in identifying equal-quality translations** while S×S RR offered a cost-effective alternative with limited ability in capturing subtle quality differences due to the lack of error annotation.[2] In terms of error distribution, **S×S MQM can affect the distribution of error categories and severities**, as seen in the higher detection rate of accuracy errors in ZhEn compared to MQM. Overall, the contributions of this paper are:

(1) It examines the point-wise MQM, S×S MQM, and the simplified S×S RR, offering a systematic investigation of comparative judgment in human annotation tasks in MT;

(2) It offers insights that S×S MQM provides more reliable and fine-grained annotations while S×S RR provides an efficient alternative when detailed error annotation is not required.

[2]We analyzed annotation time per output and found that, compared to MQM, S×S MQM takes approximately 20% more time, while S×S RR requires about 60% less. The slight increase for S×S MQM may stem from the added effort of comparing two translations while providing fine-grained annotations.

2 Annotation settings

This section introduces the core concepts and methodologies used in this study. It begins with key terminology in Section 2.1, followed by the rationale for integrating comparative judgment into annotation tasks (Section 2.2) and an overview of the MQM framework in Section 2.3. The three annotation settings analyzed in this work—MQM, S×S MQM, and S×S RR—are then described in detail in Section 2.4. Finally, Section 2.5 explains how annotations are converted into numeric scores.

2.1 Terminology

Two terms in the current work are defined here following Riley et al. (2024). A **segment** is a unit of one or multiple sentences that is highlighted at a time for annotators to focus on for annotation (i.e., the grey-highlighted text in Figure 1); a **document** is a sequence of input segments (e.g., an excerpt from an article).

2.2 Side-by-side annotation

Comparative judgment (Thustone, 1927) is a psychometric method that presents two items side by side, asking which better satisfies a given criterion. It assumes that people are more reliable when comparing two items than evaluating them individually. Studies show that comparative judgment improves the consistency and accuracy of teachers' assessment of students' writing (Pollitt, 2012; de Moira et al., 2022; Jones and Davies, 2024).

In natural language generation, the S×S annotation (more commonly called *pair-wise annotation*) has been used for tasks like open-ended text generation (Wang et al., 2023; Krishna et al., 2023), machine translation (Karpinska and Iyyer, 2023), and long-form question answering (Xu et al., 2023). However, there lacks systematic study that compares this approach with point-wise approaches in terms of annotator behavior and annotation results.

2.3 Multidimensional quality metrics

MQM is an annotation framework proposed by Lommel et al. (2014b) and refined by Freitag et al. (2021a). It is the state-of-the-art annotation setting currently used by WMT (Freitag et al., 2023). It involves marking error spans and assigning error severity and category. In this work, annotators use the error hierarchy in Table 8 (Appendix A) for error categorization. All categories can be either `major` or `minor`, except for `Non-translation` which is always `major`.[3] Annotators are instructed that the more precise the error spans, the more informative the annotation.

2.4 Studied annotation settings

Three annotation settings, illustrated in Figure 1, are studied and compared to each other: MQM, S×S MQM, and S×S RR. Annotators evaluate system outputs segment by segment, with full access to the surrounding context.

Single-sided MQM (i.e., point-wise MQM; henceforth, MQM), as in Section 2.3, involves annotators evaluating one translation at a time.

S×S MQM is the same as MQM with one difference: two translations (each from a different system) are shown side-by-side instead of just one. The core annotation task is otherwise unchanged.

S×S RR does not require detailed error annotation. Annotators view two translations of the same input text side-by-side and rate them on a five-point scale as `(much) better` on one side or `about the same` (see Figure 1).[4] This setting evaluates whether side-by-side annotation can provide reliable system rankings without detailed error annotation.

2.5 Score calculation

Expert annotations are converted into numeric scores for system and annotation setting comparison. For MQM and S×S MQM, the segment score is the average of the scores assigned by each annotator, with each score determined by the error severity and category (Table 9 in Appendix B). The system score is obtained by averaging the segment scores. For MQM, lower scores are better, with an error-free segment receiving a score of 0. The scores are z-normalized following Riley et al. (2024).

The scoring of S×S RR follows MQM in that a lower score is better. In each rated pair, the `much worse` translation segment is penalized by 2 points, and the `worse` segment by 1 point. If both translations are of similar quality, no penalty is applied. As the MQM settings, the segment score is the average of the scores assigned by each annotator. The system score is averaged over the segment scores.[5]

3 Experimental setup

This section details the setup of human experiments, including the dataset, the criteria for selecting systems to be evaluated, and the process for assigning tasks among annotators.

3.1 Dataset and language pairs

The human annotation experiments are performed on the system outputs from the news domain in WMT2023 (Freitag et al., 2023), covering two language pairs: Chinese to English (`ZhEn`) and English to German (`EnDe`). Basic statistics are in Table 1.

	ZhEn	EnDe
Documents	38	30
Segments	377	104
Avg. English tokens/seg	32.02	71.91

Table 1: Basic dataset statistics. For `ZhEn`, average tokens per segment are based on the English reference translation, and for `EnDe`, on the English source. Tokens are counted using whitespace in both cases.

3.2 System pairs

Due to the time and cost involved in human evaluation, pairwise comparisons of all MT systems in the S×S settings are impractical. Therefore, 5 system pairs are selected per language pair (Table 2), which are drawn from the systems in the WMT 2023 General Machine Translation Task (Kocmi et al., 2023).

[3] `Major` errors significantly alter the meaning of the source text; `minor` errors are noticeable but do not significantly alter the source meaning.

[4] Unlike the relative ranking used in, for example, WMT13 (Macháček and Bojar, 2013), S×S RR provides document context to help translation quality assessment. See Section 6 for further comparison.

[5] We z-normalized the (S×S) MQM scores to account for variations in score ranges across different annotators, ensuring comparability and mitigating individual annotator biases. In contrast, the S×S RR scores were not z-normalized since they are inherently constrained within the range of -2 to 0.

	System Pairs	Rank	p	Cross-BLEU	Criteria
ZhEn	GPT4-5shot Lan-BridgeMT	1 2	0.05	62.2	Top 2 systems
ZhEn	HW-TSC ONLINE-A	4 5	0.27	57.3	High text similarity
ZhEn	IOL-Research ONLINE-B	6 8	0.17	52.2	High text similarity
ZhEn	ONLINE-W NLLB_Greedy	10 12	0.10	31.2	Lower text similarity
ZhEn	NLLB_BLEU ONLINE-M	14 15	0.40	35.7	Lower text similarity
EnDe	ONLINE-W GPT4-5shot	1 2	0.09	53.1	Top 2 Systems
EnDe	ONLINE-Y ONLINE-A	5 6	0.48	62.0	High text similarity
EnDe	ONLINE-M ONLINE-G	7 8	0.22	56.3	High text similarity
EnDe	GPT4-5shot refA	2 3	0.11	39.3	Lower text similarity
EnDe	NLLB_BLEU Lan-BridgeMT	10 11	0.24	44.3	Lower text similarity

Table 2: Systems annotated in the human experiments. The ranks are determined by XCOMET. The p values are calculated by a random permutation test with 10000 trials to determine quality similarity. Cross-BLEU quantifies the text similarity of system outputs.

Two features are considered when forming system pairs: text similarity and quality similarity. For text similarity, cross-BLEU (Papineni et al., 2002) is applied. For quality similarity, XCOMET-QE-Ensemble (XCOMET) (Guerreiro et al., 2024; Freitag et al., 2023) is used in tandem with a random permutation test.[6] XCOMET provides segment scores for each system. Two systems are considered similar in quality if the permutation test of their segment scores returns $p > 0.05$.

For each language pair, the top two systems identified by XCOMET form a pair; two pairs have high text similarity (cross-BLEU); and two have low text similarity. In all cases, systems with similar quality were selected, as distinguishing small quality differences presents a greater challenge in practical scenarios. The system pairs are listed in Table 2.[7] Applying these criteria resulted in GPT4-5shot appearing twice in EnDe. This has the benefit of allowing comparison between each instance of this system's evaluation in the side-by-side settings.

3.3 Task assignment

The annotation experiments are conducted by professional translators who regularly perform MQM annotation. Tasks are distributed approximately evenly among 8 annotators for ZhEn and 10 for EnDe. To mitigate rater bias, we use the within-subject setup of Riley et al. (2024): for each input document, all system translations for that document are evaluated by the same set of 3 annotators. Additionally, each translation is evaluated by the same 3 annotators in all 3 annotation settings.

4 Meta evaluation of human evaluations

This section outlines five criteria for analyzing the human annotation results: (1) inter-annotator agreement, (2) inter-translation consistency, (3) system-level ranking, (4) segment-level ranking, and (5) error distribution. Results are in Section 5.

Inter-annotator agreement (IAA) Each segment pair is annotated by three annotators, allowing for calculating the IAA. For each segment translated by system a and b, each annotator's annotation is categorized as $a > b$, $a = b$, or $a < b$. Krippendorff's α (Krippendorff, 2018) quantifies the IAA.[8]

Inter-translation consistency When the same error occurs in translations from multiple systems of the same source input, annotators should mark it consistently with the same span, category, and severity. This consistency is crucial for fair system comparisons and training MQM-style automatic metrics (Juraska et al., 2023; Fernandes et al., 2023). Inter-translation consistency quantifies the degree to which annotators achieve this uniformity across translations. The detailed process for calculating this consistency is provided in Appendix C.2.

Agreement in segment-level rankings Pairwise ranking agreement (PRA) (Deutsch et al., 2023), defined in Equation 1 and Table 10 (Appendix C.1), measures the consistency between two annotation settings in ranking translation pairs by considering agreements, disagreements, and ties to evaluate alignment between evaluation methods.[9]

Agreement in system-level rankings Systems in each pair in Table 2 are ranked pairwise based on scores calculated as detailed in Section 2.5. A random permutation test is applied to the segment-level scores of the paired systems to evaluate the statistical significance of observed differences.

Error distribution For the two MQM settings, one possibility is that they identify more/fewer errors of particular kinds. To explore this, the total number

[6] `permutation_test` from `scipy.stats`.

[7] For brevity, NLLB_MBR_BLEU is referred to as NLLB_BLEU in this work.

[8] In the MQM settings, two segments tie if they have the same score.

[9] The metric is termed *pairwise accuracy* in Deutsch et al. (2023). However, since there is no gold reference in this work, the metric is referred to as *agreement*. The word *ranking* is to emphasize that the metric pertains to rankings.

of target-side errors annotated across all 3 ratings was counted for each language pair and MQM setting by category and severity.

5 Results and discussion

This section presents the meta-evaluation of the human experiments using the metrics outlined in Section 4, with results reported based on z-normalized scores following Riley et al. (2024). Segments annotated by a ZhEn outlier annotator are excluded.[10] The findings show that comparative judgment improves annotator agreement and consistency, maintains a reliable quality ranking, and facilitates accuracy error finding in ZhEn.

5.1 Inter-annotator agreement

S×S settings consistently yield higher agreement, particularly in S×S MQM, in both ZhEn and EnDe. Table 3 presents the Krippendorff's α, indicating fair agreement among annotators in the three annotation protocols. The IAA does not exhibit a clear correlation with textual similarity between systems, as detailed in Table 13 (Appendix E), an expanded version of Table 3.

	MQM	S×S MQM ↑	S×S RR ↑
ZhEn	0.2178	**0.2510**	0.2380
EnDe	0.2345	**0.3594**	0.2402

Table 3: Krippendorff's α in three annotation settings. The annotators in each setting achieve a fair agreement.

The results suggest that comparative judgment improves alignment among human annotators in evaluations. This is likely because MQM, as a pointwise approach, introduces more noise by preventing direct comparisons between translations. In contrast, the S×S settings allow for direct comparisons, reducing noise by minimizing inconsistent error marking (Section 5.2) and instances where shared mistakes are flagged for one system but overlooked for the other.

We hypothesize two reasons for S×S MQM's higher agreement compared to S×S RR.

First, S×S MQM enables explicit error marking, reducing ambiguity and enhances the clarity of the decision-making process. In contrast, S×S RR requires annotators to simultaneously evaluate and weigh multiple aspects of two segments (e.g., accuracy and style). This increases cognitive load and introduces greater variability in their decisions.

Second, the increased cognitive load may cause annotators to be influenced by longer segments during comparative judgments. To test this, we ranked the segments by length, divided them into three equally sized groups, and computed Krippendorff's α for each group. The results in Table 12 (Appendix E) show that, in the S×S settings, the shortest segments achieve the highest agreement.

Overall, the improvement in segment-level agreement introduced by comparative judgment is valuable because segment-level evaluation is susceptible to noise (Freitag et al., 2023), so mitigating that noise can improve reliability.

5.2 Inter-translation consistency

S×S MQM demonstrates remarkable increases in inter-translation error marking consistency, as shown in Table 4.[11] The upper half of Table 4 shows a substantial consistency increase, averaging 40.4% for ZhEn and 35.7% for EnDe, when evaluating two systems together in S×S MQM (i.e., the pairs in Table 2). This improvement persists, averaging 20.4% for ZhEn and 18.6% for EnDe, even when two systems are not evaluated side-by-side, as shown in the lower table.[12] All increases extends beyond error spans into categories and severity. The improvement in non-compared systems in S×S MQM may result from exposure to side-by-side comparisons, which potentially refine annotators' internal error detection standards as well as increase error awareness and cognitive anchoring. Lastly, the consistency gains are smaller for systems with lower text similarity (Table 15 in Appendix F), likely due to less surface overlap.

The findings demonstrate that comparative judgment significantly enhances annotators' consistency in identifying error spans and assigning error severity and categories. The improvement is valuable for both gaining insights from annotations and training MQM-style automated metrics.

[10] The rationale for excluding the outlier annotator is provided in Appendix D. After excluding the segments annotated by the outlier annotator, ZhEn has 16 documents with 220 segments and an average token count of 31.48 per segment.

[11] The results without removing the ZhEn outlier annotator are in Table 14, which remains similar to Table 4, meaning that, while the outlier annotator identified significantly more errors, they also exhibited improved inter-translation consistency.

[12] The ZhEn results are the average of 40 system pairs. For EnDe, because GPT4-5shot is annotated twice in S×S MQM, one of the annotated GPT4-5shot is excluded. Hence, the last two rows in Table 14 are the averages of 31 pairs.

	Setting	Span ↑	Span + Cat. ↑	Span + Sev. ↑	Span + Cat. + Sev. ↑
Inter-translation consistency from explicitly compared systems (5 pairs)					
ZhEn	MQM	26.78%	24.66%	26.18%	24.25%
	S×S MQM	67.50%	67.03%	67.26%	66.85%
EnDe	MQM	45.07%	41.61%	42.81%	40.25%
	S×S MQM	78.7%	77.84%	78.48%	77.67%
Inter-translation consistency from *not* explicitly compared systems (ZhEn: 40 pairs; EnDe: 31 pairs)					
ZhEn	MQM	26.77%	24.63%	26.06%	24.15%
	S×S MQM	46.90%	45.77%	46.00%	45.12%
EnDe	MQM	45.60%	42.87%	43.81%	41.60%
	S×S MQM	63.53%	62.39%	61.67%	60.69%

Table 4: Inter-translation consistency, averaged over 7 (ZhEn) and 10 (EnDe) annotators, in MQM and S×S MQM. Cat. = category, Sev. = severity. Inter-translation consistency is calculated for four criteria of what counts as common errors in two systems, for example, Span + Cat. = errors with the same span *and* category. For EnDe, the annotation of GPT4-5shot in pair with ONLINE-W is not included in the calculation of the lower table results. The green color highlights the higher values between MQM and S×S MQM.

α setting	β setting	ZhEn PRA ↑	EnDe PRA ↑	Avg.
MQM	S×S RR	0.568	0.540	0.554
S×S MQM	S×S RR	**0.626**	0.629	0.628
MQM	S×S MQM	0.623	**0.646**	0.635

Table 5: Segment pairwise ranking agreement between every two annotation settings. Results are based on z-scores with the ZhEn outlier annotator being excluded.

Language pair	MQM	S×S MQM	S×S RR
ZhEn	7.36%	16.55%	18.55%
EnDe	6.92%	11.54%	16.15%

Table 6: Tie rate in three annotation settings. S×S RR has the highest tie rate in both language pairs.

5.3 Segment-level ranking agreement

S×S MQM and S×S RR show solid agreement with each other and are better at identifying equal-quality segments. Table 5 presents the PRA results for every pair of settings in ZhEn and EnDe. Table 6 reports the tie rates for each pair of settings. The results provide three important insights.

First, MQM and S×S RR have the lowest agreement in both language pairs, largely due to the fundamental differences in their features: pointwise *vs*. pairwise and detailed error annotation *vs*. preference only. This shows that methodological divergence indeed impacts annotation outcomes.

Second, S×S MQM and S×S RR show solid agreement (Table 5). With better IAA than MQM (Table 3) and lower cost than S×S MQM, S×S RR is an appealing and efficient choice when detailed error annotation is not required.

Third, the MQM setting has the lowest tie rate (Table 6). This can be attributed to the fact that MQM lacks explicit comparisons between paired segments, which results in its low inter-translation consistency (Table 4). As a result, MQM may misjudge segment pairs of equal quality, compromising the reliability of its annotation outcomes.

5.4 System-level ranking agreement

MQM, S×S MQM, and S×S RR demonstrate strong agreement in system-level rankings; S×S RR's high tie rate may impact its reliability. Table 7 presents the system ranking results.

For ZhEn, all three settings yield identical system rankings. For EnDe, GPT4-5shot was annotated twice in S×S MQM, once with ONLINE-W and once with refA, and obtained stable scores. This suggests that S×S MQM does not compromise the absolute evaluation of individual systems. S×S RR on ONLINE-A and ONLINE-Y show a discrepancy with MQM and S×S MQM; however, the difference is not statistically significant, indicated by the p-value.

Further investigation into the discrepancy in EnDe S×S RR reveals that the high tie rate in S×S RR (Table 6) plays a key role. In three rounds of S×S RR annotations, ONLINE-Y and ONLINE-A tied in two. Across all annotations from three annotators, the tie outcome occurred 116 times (37.18%) in S×S MQM, compared to 17.31% in MQM and 29.48% with S×S MQM.

These findings suggest several insights: (1) the coarse rating scale of S×S RR makes it difficult to detect nuanced quality differences; (2) while annotating error in MQM facilitates fine-grained distinctions, it also increases the risk of spurious differences due to rater noise; and (3) S×S MQM bal-

	Setting	Better System	Worse System	p value
Top 2	MQM	Lan-BridgeMT (-0.33)	GPT4-5shot (-0.28)	**0.013**
	S×S MQM	Lan-BridgeMT (-0.26)	GPT4-5shot (-0.21)	**0.025**
	S×S RR	Lan-BridgeMT (0.36)	GPT4-5shot (0.47)	**0.013**
High text sim	MQM	HW-TSC (-0.20)	ONLINE-A (-0.18)	0.277
	S×S MQM	HW-TSC (-0.17)	ONLINE-A (-0.14)	0.234
	S×S RR	HW-TSC (0.33)	ONLINE-A (0.52)	**0.000**
	MQM	ONLINE-B (-0.13)	IOL_Research (-0.11)	0.213
	S×S MQM	ONLINE-B (-0.17)	IOL_Research (-0.10)	**0.014**
	S×S RR	ONLINE-B (0.35)	IOL_Research (0.54)	**0.000**
Low text sim	MQM	ONLINE-W (0.02)	NLLB_Greedy (0.48)	**0.000**
	S×S MQM	ONLINE-W (0.02)	NLLB_Greedy (0.41)	**0.000**
	S×S RR	ONLINE-W (0.29)	NLLB_Greedy (0.75)	**0.000**
	MQM	ONLINE-M (0.20)	NLLB_BLEU (0.50)	**0.000**
	S×S MQM	ONLINE-M (0.19)	NLLB_BLEU (0.40)	**0.000**
	S×S RR	ONLINE-M (0.31)	NLLB_BLEU (0.61)	**0.000**

(a) Chinese → English

	Setting	Better System	Worse System	p value
Top 2	MQM	ONLINE-W (-0.32)	GPT4-5shot (-0.27)	0.075
	S×S MQM	ONLINE-W (-0.35)	GPT4-5shot (-0.29)	0.070
	S×S RR	ONLINE-W (0.33)	GPT4-5shot (0.45)	**0.046**
High text sim	MQM	ONLINE-A (-0.16)	ONLINE-Y (-0.08)	**0.029**
	S×S MQM	ONLINE-A (-0.18)	ONLINE-Y (-0.10)	**0.014**
	S×S RR	ONLINE-Y (0.38)	ONLINE-A (0.43)	0.258
	MQM	ONLINE-M (0.02)	ONLINE-G (0.13)	0.058
	S×S MQM	ONLINE-M (0.08)	ONLINE-G (0.16)	0.15
	S×S RR	ONLINE-M (0.43)	ONLINE-G (0.44)	0.448
Low text sim	MQM	refA (-0.35)	GPT4-5shot (-0.27)	**0.037**
	S×S MQM	refA (-0.32)	GPT4-5shot (-0.31)	0.412
	S×S RR	refA (0.36)	GPT4-5shot (0.50)	**0.027**
	MQM	Lan-BridgeMT (0.39)	NLLB_BLEU (0.63)	**0.005**
	S×S MQM	Lan-BridgeMT (0.44)	NLLB_BLEU (0.87)	**0.000**
	S×S RR	Lan-BridgeMT (0.22)	NLLB_BLEU (0.86)	**0.000**

(b) English → German

Table 7: Pairwise system rankings and statistical significance of system quality differences for ZhEn and EnDe under three annotation settings. The results are based on the z-normalized scores with the ZhEn outlier annotator being excluded. The red highlight points out a system ranking discrepancy. The p values indicate statistical significance in system quality differences, determined by a random permutation test with 10000 trials.

ances these trade-offs more effectively, as reflected in its higher inter-translation consistency.

5.5 MQM error distribution

S×S MQM highlights more major accuracy errors in ZhEn, reflecting its ability in finding accuracy errors that may be neglected in MQM. Figure 2 illustrates the distribution of error category percentages for MQM and S×S MQM.[13]

While EnDe S×S MQM shows similar proportions to MQM, ZhEn S×S MQM shows a higher prevalence of major accuracy errors, further supported by the detailed counts in Figure 3. To understand the source of major accuracy errors in ZhEn S×S MQM, we examined whether annotators altered their category assignment of the same errors across the two annotation settings. The heatmaps in Figure 5 (Appendix G) reveal category conversions, notably from Terminology and Style errors in MQM to Accuracy errors in S×S MQM for ZhEn.

The conversions only partially explain the increase in major accuracy errors in ZhEn. A review of 50 randomly sampled segments from the ZhEn ONLINE-A system[14] revealed that many accuracy errors identified in S×S MQM were not annotated in MQM.

Overall, EnDe has more fluency errors in both MQM settings while ZhEn has a significant increase in accuracy errors in S×S MQM. This may stem from English and German belonging to the same language family, making fluency the key challenge in translation, whereas the linguistic differences between Chinese and English make accuracy a greater challenge in ZhEn. S×S MQM may further highlight accuracy errors in ZhEn, especially when only one translation contains such an error.

[13]Because GPT4-5shot is annotated twice in EnDe S×S MQM, when counting error numbers, the GPT4-5shot errors in MQM are duplicated for a fair comparison between EnDe MQM and S×S MQM.

[14]ONLINE-A shows the largest increase in accuracy errors when comparing S×S MQM to MQM.

6 Related work

Human evaluation has long been the gold standard for assessing both MT performance and automatic MT evaluation metrics in the WMT conferences and numerous MT studies. Over the years, it has evolved through efforts to establish more reliable and replicable methods (Stanchev et al., 2020).

Human evaluation in the WMT metrics shared task transitioned from the 5-point scale on fluency and adequacy (Koehn and Monz, 2006) to relative ranking (RR) of 5 translation sentences or phrases (Callison-Burch et al., 2007) for a better inter-/intra-annotator agreement (Callison-Burch et al., 2007, 2008). WMT16 (Bojar et al., 2016) used direct assessment (DA) (Graham et al., 2013, 2016, 2017), where annotators scored translations between 0 and 100. It returns reliable evaluation when each item receives 15 or more judgments. As MT systems advance, DA struggles and mistakenly ranks high-quality human translations below machine outputs (Freitag et al., 2021b). To improve human evaluation quality, MQM (Lommel et al., 2014b; Freitag et al., 2021a) was introduced into WMT21 (Freitag et al., 2021b). It emphasizes the inclusion of context (Mathur et al., 2020) and the use of experts to better capture subtle differences (Goto et al., 2014; Toral et al., 2018; Läubli et al., 2020).

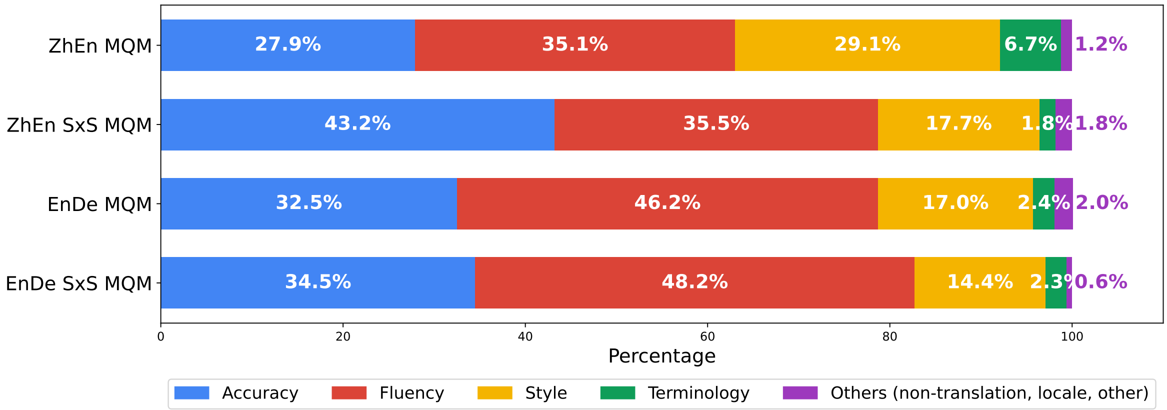


Figure 2: Percentages of error categories in the MQM settings in ZhEn and EnDe. The GPT4-5shot errors in EnDe MQM are doubled for a fair comparison with EnDe S×S MQM. While the percentages in EnDe stay relatively stable, in ZhEn, accuracy errors have a higher percentage in S×S MQM than in MQM.

MQM is the state-of-the-art MT evaluation method, but it has shortcomings—disagreement in marking error span boundaries, category, and severity (Lommel et al., 2014a). By introducing S×S MQM, the current work aims to address those issues. Kocmi et al. (2024) aim to mitigate the impact of these issues through Error Span Annotation (ESA), a point-wise annotation setting where annotators first identify error spans (with severity) in a segment before assigning it an overall score. The segment-level scores are different from MQM in two ways: first, the scores are assigned to measure the amount of meaning preserved in translations (as in their Figure 1); second, the scores are not automatically calculated from errors, which may introduce subjectivity and latency. For direct system comparisons, S×S RR and S×S MQM may be more effective, as comparative judgments are cognitively easier and more consistent than absolute ratings (Thustone, 1927), with differences between outputs naturally highlighted. A promising future direction is adapting ESA to a pairwise format, although this requires careful design, for instance, ensuring score comparability and consistency within and across pairs.

Pairwise evaluation offers a simpler and more intuitive approach to MT evaluation. Previous WMT workshops ranked system outputs using 5-way relative ranking, which was then converted into pairwise comparisons (Bojar et al., 2013, 2014). Vilar et al. (2007) advocate for using binary instead of n-ary RR, as it is more intuitive and straightforward for annotators. Unlike S×S RR, their method does not provide context and requires annotators to rank segments based on only adequacy and fluency. All possible system pairs are considered, with a full system ranking being obtained either by treating the task as a sorting problem or by applying the Bradley-Terry model (Bradley and Terry, 1952; Dras, 2015). Pairwise evaluation is also implemented in automatic MT evaluation, for example, by Guzmán et al. (2015) and Liu et al. (2024).

7 Conclusion

This study uses machine translation as a case study and examines the impact of MQM, S×S MQM, and S×S RR on the annotation results from five aspects: inter-annotator agreement, inter-translation error annotation consistency, quality ranking at segment- and system-levels, and error distributions.

Incorporating comparative judgment, S×S MQM and S×S RR achieved higher inter-annotator agreement. S×S MQM enhanced error marking consistency both for explicitly compared system pairs and across others. Concerning S×S RR, although it does not provide detailed error annotations, it offers an efficient and reliable alternative for system ranking provided by S×S MQM, with a trade-off in differentiating subtle quality differences due to higher tie rates.

The findings in the paper demonstrate the value of comparative judgment in improving annotation quality and efficiency, with S×S MQM and S×S RR serving as practical alternatives to MQM, tailored to different evaluation needs.

Limitations

Practical considerations make it difficult to control all variables in human evaluation experiments. As an example, the single-sided MQM annotations

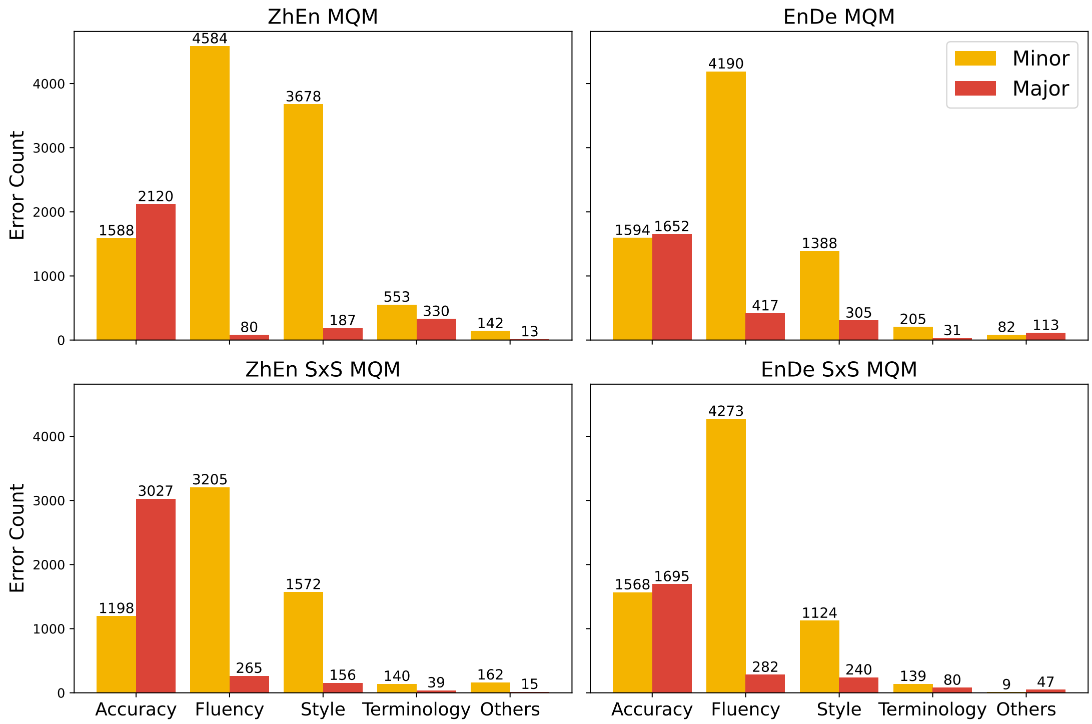


Figure 3: Number of errors in five categories in the MQM settings in `ZhEn` and `EnDe` of all three rounds of annotations. `Others` includes non-translation, locale convention, and other.

were collected in 2023 with a different original goal, while the S×S MQM and S×S RR annotations were collected in 2024 for this project. Additionally, our annotators were engaged in multiple projects throughout 2024 and thus may have performed other annotation tasks in between items collected for this project.

Due to the time and cost involved in human evaluation, the current work is not able to test MQM, S×S MQM, and S×S RR on a larger set of language pairs besides `ZhEn` and `EnDe`. It is also impractical to exhaustively test all possible system pairs in the two tested language pairs. However, the current study still provides a strong foundation for understanding the trade-offs between detailed error detection and overall system ranking. Future work could expand this work to include more language pairs and a broader range of systems, further validating the generalizability of the results.

We focus on pairwise side-by-side annotation, as its simplicity helps annotators more easily compare translation quality. While comparing more than two outputs is possible (Macháček and Bojar, 2013), it increases cognitive load and may reduce agreement, especially in setups without explicit error marking. We believe this is an important direction that warrants dedicated study.

Beyond MT, future work can test the pairwise annotation setups in other domains to see how annotation settings influence evaluations in diverse NLP tasks. Additionally, researchers can investigate the impact of different annotation settings on annotator backgrounds. Although S×S MQM and S×S RR do not significantly increase the inter-annotator agreement of expert annotators, they might do so for crowd-sourced workers.

Ethics statement

Our professional translator annotators were sourced by a translation agency. They were given sufficient time to complete the task and paid fairly.

The annotators worked on the news data from WMT2023 (Freitag et al., 2023) that should not contain offensive content. The annotation task did not ask for personally identifiable information.

Acknowledgements

We thank the MQM raters for their contributions to our dataset. We are also grateful to Eleftheria Briakou for her thoughtful feedback on an early draft. We appreciate the anonymous reviewers for their comments, which helped improve the final version of the paper. Finally, we thank Marzena Karpinska and Kalpesh Krishna for their helpful discussions and support.

References

Anja Belz and Eric Kow. 2010. Comparing rating scales and preference judgements in language evaluation. In *Proceedings of the 6th International Natural Language Generation Conference*. Association for Computational Linguistics.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria. Association for Computational Linguistics.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Ondřej Bojar, Yvette Graham, Amir Kamran, and Miloš Stanojević. 2016. Results of the WMT16 metrics shared task. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 199–231, Berlin, Germany. Association for Computational Linguistics.

Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39:324.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic. Association for Computational Linguistics.

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

Anne Pinot de Moira, Christopher Wheadon, and Daisy Christodoulou. 2022. The classification accuracy and consistency of comparative judgement of writing compared to rubric-based teacher assessment. *Research in Education*, 113(1):25–40.

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.

Mark Dras. 2015. Squibs: Evaluating human pairwise preference judgments. *Computational Linguistics*, 41(2):309–317.

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, 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, 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.

Shinsuke Goto, Donghui Lin, and Toru Ishida. 2014. Crowdsourcing for evaluating machine translation quality. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3456–3463, Reykjavik, Iceland. European Language Resources Association (ELRA).

Yvette Graham, Timothy Baldwin, Meghan Dowling, Maria Eskevich, Teresa Lynn, and Lamia Tounsi. 2016. Is all that glitters in machine translation quality estimation really gold? In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3124–3134, Osaka, Japan. The COLING 2016 Organizing Committee.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2017. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1):3–30.

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

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

Ian Jones and Ben Davies. 2024. Comparative judgement in education research. *International Journal of Research & Method in Education*, 47(2):170–181.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics.

Marzena Karpinska, Nader Akoury, and Mohit Iyyer. 2021. The perils of using Mechanical Turk to evaluate open-ended text generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1265–1285, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. In *Proceedings of the Eighth Conference on Machine Translation*, pages 419–451, Singapore. 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, Maja Popović, and Mariya Shmatova. 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, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024. Error span annotation: A balanced approach for human evaluation of machine translation. *Preprint*, arXiv:2406.11580.

Philipp Koehn and Christof Monz. 2006. Manual and automatic evaluation of machine translation between european languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121, New York City. Association for Computational Linguistics.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo. 2023. LongEval: Guidelines for human evaluation of faithfulness in long-form summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1650–1669, Dubrovnik, Croatia. Association for Computational Linguistics.

Samuel Läubli, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral. 2020. A set of recommendations for assessing human-machine parity in language translation. *J. Artif. Intell. Res.*, 67:653–672.

Yinhong Liu, Han Zhou, Zhijiang Guo, Ehsan Shareghi, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2024. Aligning with human judgement: The role of pairwise preference in large language model evaluators. In *First Conference on Language Modeling*.

Arle Lommel, Maja Popovic, and Aljoscha Burchardt. 2014a. Assessing inter-annotator agreement for translation error annotation. In *MTE: Workshop on Automatic and Manual Metrics for Operational Translation Evaluation*, pages 31–37. Language Resources and Evaluation Conference Reykjavik.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014b. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica*, pages 0455–463.

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

Matouš Macháček and Ondřej Bojar. 2013. Results of the WMT13 metrics shared task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 45–51, Sofia, Bulgaria. Association for Computational Linguistics.

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.

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

Chau Minh Pham, Simeng Sun, and Mohit Iyyer. 2024. Suri: Multi-constraint instruction following for long-form text generation. *Preprint*, arXiv:2406.19371.

Alastair Pollitt. 2012. The method of adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice*, 19(3):281–300.

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.

Peter Stanchev, Weiyue Wang, and Hermann Ney. 2020. Towards a better evaluation of metrics for machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 928–933, Online. Association for Computational Linguistics.

Louis Leon Thustone. 1927. A law of comparative judgement. *Pschological Review*, 34:273–286.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Brussels, Belgium. Association for Computational Linguistics.

David Vilar, Gregor Leusch, Hermann Ney, and Rafael E. Banchs. 2007. Human evaluation of machine translation through binary system comparisons. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 96–103, Prague, Czech Republic. Association for Computational Linguistics.

Shufan Wang, Yixiao Song, Andrew Drozdov, Aparna Garimella, Varun Manjunatha, and Mohit Iyyer. 2023. kNN-LM does not improve open-ended text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15023–15037, Singapore. Association for Computational Linguistics.

Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. 2023. A critical evaluation of evaluations for long-form question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3225–3245, Toronto, Canada. Association for Computational Linguistics.

A MQM error categories

Table 8 lists all the error categories and their subcategories mentioned in Section 2.3. The annotators used the (sub)categories to label errors in the MQM and S×S MQM settings.

Category	Subcategory
`Accuracy`	Reinterpretation Mistranslation Gender Mismatch Untranslated Addition Omission
`Fluency`	Inconsistency Grammar Register Spelling Text-Breaking Punctuation Character Encoding
`Style`	Unnatural or Awkward Bad Sentence Structure Archaic or Obscure Word Choice
`Terminology`	Inappropriate for Context Inconsistent
`Locale Convention`	Address Format Date Format Currency Format Telephone Format Time Format Name Format
`Non-Translation`	—
`Other`	—
`Source Issue`	—

Table 8: MQM error hierarchy. Error spans are categorized into categories and subcategories.

B Annotation scoring scheme

Table 9 gives the scoring scheme in Section 2.5 of the MQM settings.

Severity	Category	Weight
`major`	Non-translation Others	25 5
`minor`	Fluency/Punctuation Others	0.1 1

Table 9: MQM error span weighting scheme. Gibberish segments score 25 points, `major` errors 5 points, and `minor` errors 1 point, except for `minor` punctuation errors, which are weighted at 0.1 point each.

C More details to metrics of meta evaluation

C.1 Pairwise ranking agreement

Equation 1 is used to calcualate agreement in segment-level rankings between the annotation settings and Table 10 defines its terms. The ranking of two translations of the same source segment depends on the segment scores: segments with identical scores are tied, while differing scores dictate their ranking. PRA quantifies the frequency with which two evaluation settings agree on the ranking of each pair of segments from two systems.

$$PRA = \frac{C + T_{\alpha\beta}}{C + D + T_{\alpha} + T_{\beta} + T_{\alpha\beta}} \tag{1}$$

Symbol	Description
α	One of the annotation settings
β	One of the annotation settings that is not α
C	The number of concordant pairs
D	The number of discordant pairs
T_{α}	The number of pairs tied *only* in α
T_{β}	The number of pairs tied *only* in β
$T_{\alpha\beta}$	The number of pairs tied in both α and β

Table 10: Terms in Equation 1. The annotation settings are MQM, S×S MQM, and S×S RR.

C.2 Inter-translation consistency

Using the example below, inter-translation consistency is meant to be the following: if an annotator labels "arabica" as a minor fluency error in (a), they should do the same in (b).[15] For reasons of practicality and clarity, two systems are considered at a time for calculating the consistency.

(a) Brazil is the world's largest producer of **arabica** beans, a coffee variety commonly used by baristas to make coffee.

(b) Brazil is the world's largest producer of **arabica** beans, which are the coffee beans commonly used by baristas in making coffee.

Inter-translation consistency is calculated as follows: **Alignment of Tokens** the translations from two systems are tokenized, and the alignment between their tokens is computed using `get_opcodes()` from `difflib`. This generates a list of operations (replace, delete, insert, equal) that align the tokens of the two translations; **Identification of Potential Common Errors** Errors annotated in each translation are compared based on their spans. If an error in one translation aligns with an "equal" operation in the token alignment, it is considered as a potential common error of the two compared translations. These errors are stored for further analysis; **Matching Errors Using a Criterion** A specified criterion (e.g., matching spans, categories, severities, or combinations thereof) is applied to the potential common errors. This determines how many errors are consistently marked across the two translations; **Calculation of Consistency** the consistency value is calculated as the percentage of errors that satisfy the criterion out of the total potential errors.

The final inter-translation consistency is averaged over all raters per criterion.

[15]The two translations share common spans highlighted in green, identified using the `get_opcodes()` function from Python's `difflib` module.

D Identifying and Addressing Outlier Annotators

One annotator is excluded from the `ZhEn` annotation results due to significant deviations in annotation behavior compared to their peers, which could skew the results and compromise the reliability. This decision was based on two key observations.

First, this annotator identified an exceptionally high number of errors compared to their peers. Given the mean and the standard deviation of the error counts in Table 11, the outlier annotator's z-score for `ZhEn` S×S MQM is 2.28, placing them more than two standard deviations above the mean. In contrast, all other annotators, regardless of whether they contributed to `ZhEn` or `EnDe`, remain within two standard deviations of the mean.

MQM Mean	MQM Std	S×S MQM Mean	S×S MQM Std
Chinese → English			
2936.1	1021.1	2642.3	1874.1
English → German			
900.2	413.0	945.7	475.2

Table 11: Mean and standard deviation of the error counts from 8 `ZhEn` and 10 `EnDe` annotators.

Second, visualizing the segment scores contributed by each annotator using violin plots revealed that the outlier annotator from `ZhEn` was the only one with a distinctly unimodal distribution in `ZhEn` S×S MQM, as illustrated in Figure 4. This is stark contrast to the more varied distributions observed among other annotators.

Based on the two key observations of the outlier annotator's behavior in S×S MQM, we opted to ex-

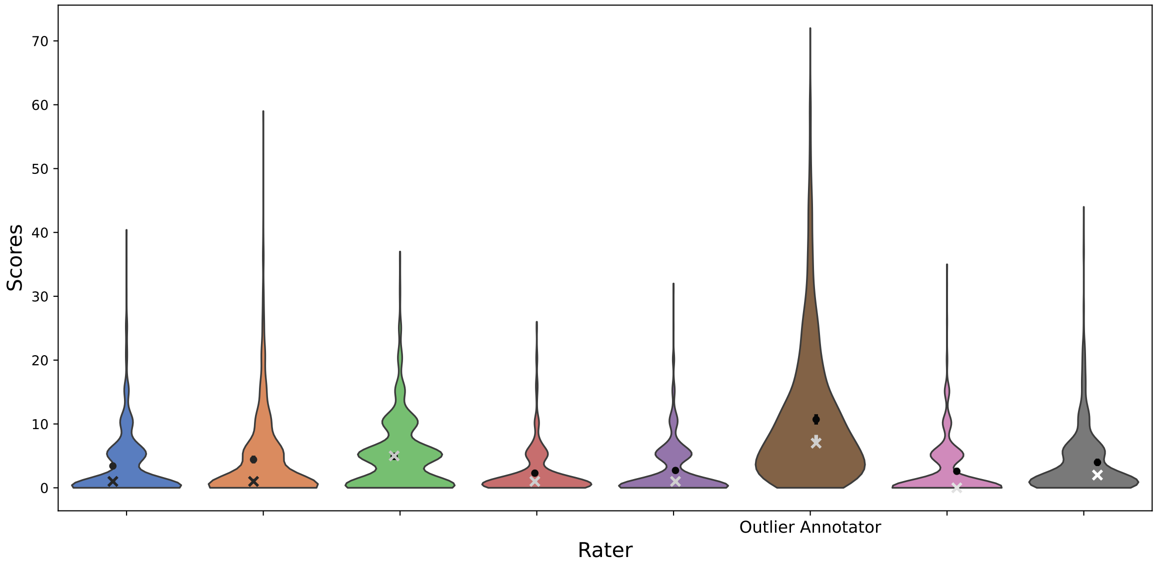


Figure 4: Violin plots of the original segment scores contributed by each annotator (without z-normalization). Annotator identities are omitted for anonymity. The dots indicate the mean, while the crosses represent the median of each distribution.

clude their annotated data from the analyses. After removing the segments annotated by the outlier annotator, there are 16 documents with 220 segments in ZhEn with an average token count of 31.45 per segment.

E Inter-annotator agreement by pair characteristics

Table 12 supports the findings made in Section 5.1 that the annotators have a higher agreement on the shortest segments in the S×S setting but the correlation between agreement levels and the segment length is not proportional. The longest segment buckets do not necessarily have the lowest α.

	Chinese → English		
Avg. toks (# seg)	12.95 (73)	27.84 (73)	53.36 (74)
MQM	**0.2183**	0.1739	0.2021
S×S MQM	**0.3154**	0.2187	0.1708
S×S RR	**0.2970**	0.1790	0.2084
	English → German		
Avg. toks (# seg)	10.62 (34)	72.94 (34)	128.83 (36)
MQM	0.1930	0.0956	**0.2307**
S×S MQM	**0.3608**	0.2763	0.2849
S×S RR	**0.3136**	0.2330	0.1645

Table 12: Krippendorff's α values for annotators' agreement across three equally sized segment groups of different average token lengths. Shorter segments elicit higher agreement in the S×S settings.

Table 13 expands Table 3 and presents the Krippendorff's α of each type of pairs (i.e., top-two, high textual similarity, and low textual similarity). In both language pairs, when evaluating the top 2 systems, the side-by-side settings achieve higher α than MQM, with S×S MQMachieving the highest. The ranking pattern for the 3 settings on the high and low text similarity system pairs is inconsistent between language pairs. For EnDe, S×S MQMachieves the highest average α in each group of system pairs.

	All	**Top 2**	**High text sim**	**Low text sim**
		Chinese → English		
MQM	0.2178	0.1560	0.2118	0.2359
S×S MQM	0.2510	0.2406	0.2290	0.2345
S×S RR	0.2380	0.2056	0.2481	0.2336
		English → Chinese		
MQM	0.2345	0.1152	0.2819	0.2213
S×S MQM	0.3594	0.2644	0.4244	0.2947
S×S RR	0.2402	0.1604	0.2189	0.2636

Table 13: Krippendorff's α of annotators' agreement in three annotation settings based on all system pairs and system pairs of three characteristics.

F Inter-translation consistency

Table 14 presents the inter-translation consistency results for ZhEn without removing the outlier annotator. The findings remain consistent with those in Table 14, showing that the outlier annotator also has a higher inter-translation consistency when doing the task using S×S MQM.

Table 15 shows that system pairs that have lower text similarity have lower consistency gains in inter-translation error marking. This is likely due to their less overlap in surface forms.

	Setting	Span ↑	Span + Cat. ↑	Span + Sev. ↑	Span + Cat. + Sev. ↑
Inter-translation consistency from explicitly compared systems (5 pairs)					
ZhEn	MQM	28.65%	26.51%	27.95%	25.99%
	S×S MQM	68.01%	67.54%	67.78%	67.35%
EnDe	MQM	45.07%	41.61%	42.81%	40.25%
	S×S MQM	78.7%	77.84%	78.48%	77.67%
Inter-translation consistency from *not* explicitly compared systems (ZhEn: 40 pairs; EnDe: 31 pairs)					
ZhEn	MQM	27.84%	25.68%	27.05%	25.11%
	S×S MQM	47.81%	46.70%	46.92%	46.04%
EnDe	MQM	45.60%	42.87%	43.81%	41.60%
	S×S MQM	63.53%	62.39%	61.67%	60.69%

Table 14: Inter-translation consistency, averaged over 8 (ZhEn) and 10 (EnDe) annotators, in MQM and S×S MQM. No annotator's annotation is removed. Cat. = category, Sev. = severity. Inter-translation consistency is calculated for four criteria of what counts as common errors in two systems, for example, Span + Cat. = errors with the same span *and* category. For EnDe, the annotation of GPT4-5shot in pair with ONLINE-W is not included in the calculation of the lower table results. The green color highlights the higher values between MQM and S×S MQM.

(a) ZhEn

Pair Type	Category	Span	Span+Cat.	Span+Sev.	Span+Cat.+Sev.
All systems	MQM	26.78	24.66	26.18	24.25
	SxS MQM	67.50	67.03	67.26	66.85
Top 2	MQM	26.55	24.62	26.48	24.55
	SxS MQM	76.46	76.46	76.46	76.46
High sim	MQM	28.56	26.45	27.97	26.08
	SxS MQM	69.13	68.35	68.72	68.06
Low sim	MQM	24.11	21.73	23.12	21.04
	SxS MQM	55.47	55.12	55.31	54.97

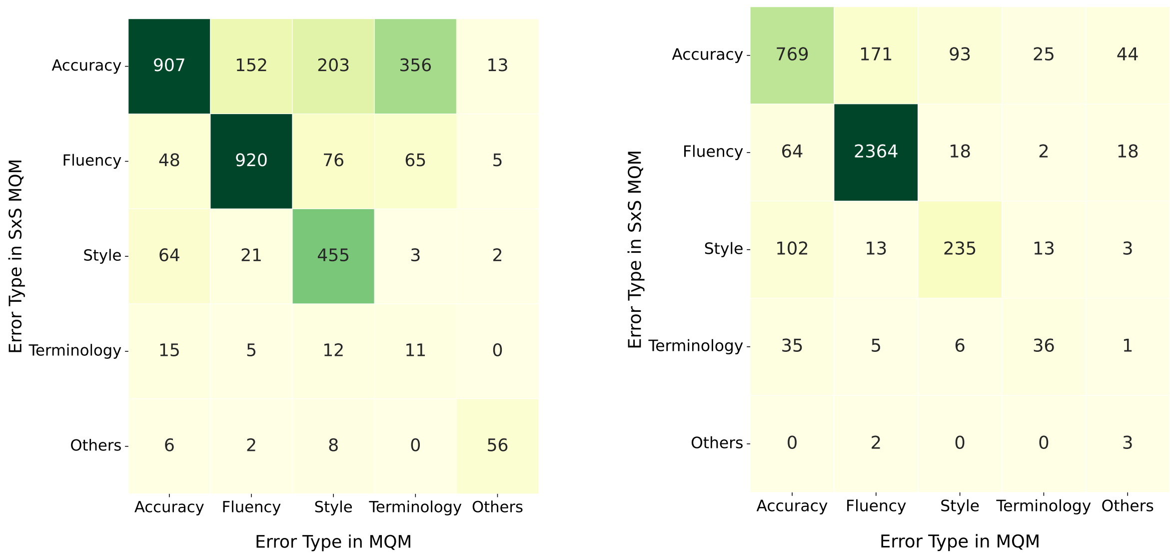
(b) EnDe

Pair Type	Category	Span	Span+Cat.	Span+Sev.	Span+Cat.+Sev.
All systems	MQM	45.07	41.61	42.81	40.25
	SxS MQM	78.70	77.84	78.48	77.67
Top 2	MQM	58.48	57.37	58.11	57.17
	SxS MQM	87.24	86.12	87.24	86.12
High sim	MQM	32.54	27.22	28.95	25.07
	SxS MQM	72.90	71.69	72.79	71.69
Low sim	MQM	45.73	42.61	43.38	41.02
	SxS MQM	72.92	72.28	72.47	71.83

Table 15: Inter-translation consistency on (a) ZhEn and (b) EnDe grouped by system pair types (see Table 2).

G Error distribution and category conversion

The heatmaps in Figure 5 demonstrate that there are some error category conversion between MQM and S×S MQM, which is more prominent in ZhEn than in EnDe.



(a) ZhEn error category conversion

(b) EnDe error category conversion

Figure 5: Error category conversion from MQM to S×S MQM in (a) ZhEn and (b) EnDe of the same errors annotated by the annotators in both MQM and S×S MQM. MQM EnDe GPT4-5shot is duplicated for the comparison.