Feeding Two Birds or Favoring One? Adequacy–Fluency Tradeoffs in Evaluation and Meta-Evaluation of Machine Translation

Behzad Shayegh[1,*] **Jan-Thorsten Peter**[2] **David Vilar**[2] **Tobias Domhan**[2]
Juraj Juraska[2] **Markus Freitag**[2] **Lili Mou**[1,3]
[1]Dept. Computing Science, Alberta Machine Intelligence Institute (Amii), University of Alberta
[2]Google [3]Canada CIFAR AI Chair, Amii
{the.shayegh, doublepower.mou}@gmail.com
{vilar, jtp, domhant, jjuraska, freitag}@google.com

Abstract

We investigate the tradeoff between adequacy and fluency in machine translation. We show the severity of this tradeoff at the evaluation level and analyze where popular metrics fall within it. Essentially, current metrics generally lean toward adequacy, meaning that their scores correlate more strongly with the adequacy of translations than with fluency. More importantly, we find that this tradeoff also persists at the meta-evaluation level, and that the standard WMT meta-evaluation favors adequacy-oriented metrics over fluency-oriented ones. We show that this bias is partially attributed to the composition of the systems included in the meta-evaluation datasets. To control this bias, we propose a method that synthesizes translation systems in meta-evaluation. Our findings highlight the importance of understanding this tradeoff in meta-evaluation and its impact on metric rankings.

1 Introduction

As translation systems become more sophisticated and widely adopted, the critical challenge of accurately evaluating their performance has come to the forefront, driving significant ongoing research to the developement of more accurate and robust metrics (Chatzikoumi, 2020; Guerreiro et al., 2024). Traditionally, translation evaluation relied on lexical metrics such as BLEU (Papineni et al., 2002) and ChrF (Popović, 2015), which primarily consider the n-gram overlap between the candidate and reference translations. However, these metrics have been shown insufficient for measuring the quality of modern translation systems (Babych and Hartley, 2008; Freitag et al., 2022). In the recent decade, researchers have been exploring the idea of training neural models to measure translation quality (Wieting et al., 2019; Ma et al., 2019; Freitag et al., 2022; Guerreiro et al., 2024), with popular examples including MetricX (Juraska et al., 2023, 2024) and Comet (Rei et al., 2020, 2022a).

Two key aspects of translation quality, long discussed in the community (Pierce and Carroll, 1966; White and O'Connell, 1993; Banchs et al., 2015; Martindale and Carpuat, 2018), are[1]:

- *Fluency*: the grammatical correctness and naturalness of the translation; and
- *Adequacy*: how well the translation conveys the meaning of the source text.

Flamich et al. (2025) demonstrate that an adequacy–fluency tradeoff exists in translation: optimizing for one aspect eventually sacrifices the other. They further introduce measurements to study the severity of this tradeoff at **the level of translation systems**.

The aforementioned tradeoff naturally results in an adequacy–fluency tradeoff at **the level of evaluation metrics**, if we limit our evaluation to a single score: increasing the capability of measuring one aspect eventually comes at the cost of decreasing the capability of measuring the other. It is important to understand this tradeoff and know the position of each metric to avoid undesired biases when optimizing translation systems according to the metric.

In this work, we demonstrate the severity of the adequacy–fluency tradeoff at the evaluation level. We analyze current evaluation setups in WMT (Callison-Burch et al., 2007; Moghe et al., 2025), and show that there is a significant disagreement between adequacy and fluency when ranking systems (Table 1). This directly imposes a severe tradeoff, as a metric can only agree with either adequacy or fluency when comparing discordant system pairs. Subsequently, we empirically analyze several contemporary translation metrics to illustrate their positions within this tradeoff.

More importantly, we show that this adequacy–

[*]Work partially done when the first author was interning at Google.

[1]Different namings and slightly different definitions are used in the literature for these two aspects; however, the main idea remains consistent.

Proceedings of the Tenth Workshop on Machine Translation (WMT), pages 269–285
November 8–9, 2025.

Evaluation Set	Concordance	Discordance
En–De'23	49 (74%)	17 (26%)
Zh–En'23	70 (67%)	35 (33%)
En–De'24	98 (72%)	38 (28%)
En–Es'24	55 (71%)	23 (29%)
Ja–Zh'24	58 (74%)	20 (26%)

Table 1: Concordance and discordance between adequacy and fluency in MQM datasets (§3.1), reported as system pair counts and percentages. Concordance means one system in a pair outperforms the other in both metrics; discordance indicates inconsistent performance across metrics. Notice that 26–33% discordance, although being the minority, is way far from being negligible and imposes a significant tradeoff when the metric quality improves.

fluency tradeoff extends even to **the level of meta-evaluation** (the evaluation of evaluation metrics). Meta-evaluation typically compares the scores of a set of candidates given by a metric to those given by humans (Callison-Burch et al., 2007). We show that the selection of the candidates used during the meta-evaluation significantly influences the identified optimal metric within the tradeoff: if there is a higher variance in the candidates' adequacy than in their fluency, the meta-evaluation will lean toward favoring metrics that prioritize adequacy. Conversely, if the fluency of these candidates shows higher variance, the meta-evaluation will prefer fluency-oriented metrics. This can inadvertently guide the development of evaluation metrics, and thus the development of translation systems, toward a particular bias toward either adequacy or fluency. We propose to synthesize candidates with desired variance in their adequacy and fluency scores to conduct a controlled (balanced) meta-evaluation.

2 Background & Related Work

2.1 Translation Metrics

The evaluation of translation systems is a critical aspect of their development and deployment, and has been researched for years (Tao et al., 2018; Chatzikoumi, 2020). This research has resulted in the development of a variety of metrics (Chauhan and Daniel, 2023; Guerreiro et al., 2024).

Traditional metrics, such as BLEU (Papineni et al., 2002), METEOR (Lavie and Denkowski, 2009), and ChrF (Popović, 2015), compare the output of a translation system against one or more human-created reference translations, only at the surface level (e.g., based on string overlap). While widely adopted for their simplicity, these surface-level metrics are shown to be incapable of measuring the quality of modern, high-quality translation systems (Freitag et al., 2022).

Recent advancements have led to the development of *trained metrics*, which leverage machine learning models to assess translation quality (Wieting et al., 2019; Freitag et al., 2022). These metrics, such as MetricX (Juraska et al., 2023, 2024) and Comet (Rei et al., 2020, 2022a), are often trained on large datasets of human annotations, allowing them to learn more nuanced correlations between system outputs and perceived quality.

Translation metrics can be divided into two categories: *reference-based* and *reference-free*. Reference-based metrics, such as those mentioned previously, rely on access to a trusted reference translation and compare the candidate against the reference to assess quality. Reference-free metrics, also known as Quality Estimation (QE) metrics, directly evaluate the candidate given the source text, without requiring a reference translation (Ito et al., 2025). This capability is particularly valuable in scenarios where human references are scarce or impractical to obtain, offering a more flexible and potentially more accurate evaluation paradigm for contemporary translation systems (Agrawal et al., 2021). MetricX and Comet offer their reference-free versions: MetricX QE (Juraska et al., 2023, 2024) and CometKiwi (Rei et al., 2022b).

In our work, we empirically analyze several contemporary translation metrics to illustrate their positions within the adequacy–fluency tradeoff, providing insights into their strengths and limitations.

2.2 Meta-Evaluation

Meta-evaluation assesses how well a translation evaluation metric correlates with human annotations (Callison-Burch et al., 2007; Macháček and Bojar, 2013). This is traditionally accomplished by computing Pearson, Kendall, and Spearman correlation coefficients (Macháček and Bojar, 2013; Mathur et al., 2020; Freitag et al., 2023), at either the segment[2] level (for individual translation scores) or the system level (for scores averaged per system). More recently, pairwise accuracy (PA; Kocmi et al., 2021) and soft pairwise accuracy (SPA; Thompson et al., 2024) have gained prominence as they directly assess how well a met-

[2]Following the common terminology in the community, we refer to each data sample as a *segment*.

ric aligns with human preferences over pairs of translations (Mathur et al., 2020). Given their popularity, we utilize these two metrics for our analysis and elaborate their formulations below.

Pairwise accuracy. PA assesses whether a metric correctly identifies the better translation (segment level) or the better system (system level) given a pair. Let m and h be the metric and human scores, respectively. PA is formulated as

$$\frac{1}{|\mathcal{P}|} \sum_{i,j \in \mathcal{P}} \mathbb{1}[\text{sgn}(m_j - m_i) = \text{sgn}(h_j - h_i)] \quad (1)$$

where $\mathcal{P}$ is the set of pairs, and sgn is the sign function. Pairs with tied human scores are excluded (Kocmi et al., 2021).

A key challenge for PA at the segment level is to handle tied metric scores. While techniques like tie calibration are employed to address this issue (Deutsch et al., 2023), recent studies (Perrella et al., 2024) indicate that such approaches are not reliable and the matter warrants further investigation. This challenge is minor at the system level, since averaging scores over many translations rarely yields ties. We therefore focus on system-level PA in our work.

Soft pairwise accuracy. SPA is a system-level meta-metric that extends PA by incorporating statistical significance from both human and metric scores. Unlike PA's binary agreement, SPA compares p-values from hypothesis tests to determine if one system is statistically superior. The SPA score is formulated as

$$\frac{1}{|\mathcal{P}|} \sum_{i,j \in \mathcal{P}} (1 - |p(h_j > h_i) - p(m_j > m_i)|) \quad (2)$$

Here, $p(x_j > x_i)$ is the p-value from a permutation test (Welch, 1990) assessing if system j is superior to system i based on scores x. SPA is sensitive to the magnitude of the score differences, reflecting the metric's confidence in its preference, and compares this confidence against that of human judgments.

3 Experimental Setup

3.1 Dataset

The Multidimensional Quality Metrics (MQM) framework provides a standardized and granular guideline for characterizing and classifying errors in translations (Mariana et al., 2015). Instead of being a score-based annotation, the MQM framework annotates translation errors. Each error is annotated as a substring of the translation with a category (such as accuracy, fluency, terminology, and style) and a severity level (usually major, minor, and neutral). This framework enables a diagnostic understanding of translation system performance, as well as more reliable and standardized evaluation scores.

WMT adopts MQM annotations (Freitag et al., 2021a) for human evaluation, producing datasets of detailed translation error annotations. We use this dataset of years 2023 and 2024 in our meta-evaluation and analysis. We reserve data from prior years (2020–2022) for training to avoid leakage. Our setup is consistent with practice in previous studies (Juraska et al., 2024; Rei et al., 2022a) for fair comparison.

It is common to transform MQM annotations into a single score (Freitag et al., 2022, 2023). This involves assigning a specific penalty to each error based on its category and severity (Freitag et al., 2021b). These penalties are then summed across all errors in the translation. We refer to this aggregated value as the *All MQM* score, where a lower score indicates higher quality.

For a more granular analysis, Flamich et al. (2025) categorize MQM errors into adequacy and fluency classes (Tables 8 and 9 in Appendix A), resulting in *Adequacy MQM* and *Fluency MQM* scores for each segment. We use these specialized scores for our tradeoff analysis. We present detailed statistics for our dataset in Appendix B.

3.2 Meta-Metrics

To study the performance of evaluation metrics in the adequacy–fluency tradeoff, we separately assess their ability to measure translation adequacy and fluency. To achieve this, we compute meta-metrics—namely, PA and SPA (§2.2)—with respect to Adequacy MQM and Fluency MQM. In Appendix C, we point out the necessity of including both PA and SPA in such studies due to their potentially different behavior in measuring bias toward adequacy or fluency.

3.3 Translation Metrics in Our Analysis

In our work, we select a set of diverse and widely used metrics to analyze their positions within the adequacy–fluency tradeoff. These metrics include **BLEU** (Papineni et al., 2002) and **ChrF** (Popović,

2015), representing overlap-based metrics; **MetricX** (Juraska et al., 2024) and **Comet** (Rei et al., 2022a), representing trained metrics; and **MetricX QE** and **CometKiwi** (Rei et al., 2023), representing reference-free metrics.

Moreover, we introduce a set of extreme translation metrics designed to measure only adequacy or fluency, which will provide a clearer view of the tradeoff between the two. Specifically, we develop **AdequacyX** and its reference-free variant, **AdequacyX QE**, trained exclusively on MQM errors categorized under adequacy. Likewise, we develop **FluencyX** trained on fluency annotations.

Technically, these trainable metrics adopt the same neural architecture as MetricX (Juraska et al., 2024), and are fine-tuned from an mT5 checkpoint (Xue et al., 2021). Following Juraska et al. (2024), we train these models on MQM annotations (WMT 2022), covering En–De, En–Ru, and En–Zh.[3] AdequacyX and AdequacyX QE are trained to predict Adequacy MQM, while FluencyX is trained to predict Fluency MQM. Unlike Juraska et al. (2024), we exclusively use MQM annotations for fine-tuning, omitting direct assessments (DA) and synthetic data, as these do not offer a clear adequacy–fluency distinction.

It should be noted that AdequacyX QE does not take the reference translation as input as it is a reference-free metric. We further restrict FluencyX to the candidate translation alone (no source or reference), similar to SENTINEL$_{\text{CAND}}$ in Perrella et al. (2024), given that fluency is considered source-independent.

To have a fair comparison, we also train our own MetricX and MetricX QE variants; they are identical to AdequacyX (QE), except that they predict All MQM scores. The key difference between our variants and the original MetricX (QE) is the exclusion of DA and synthetic data.

Additionally, we include the log-perplexities of a pretrained Gemma model as a fluency measure, following Flamich et al. (2025). We use **Gemma 3 4B** (Gemma Team, 2025) in this work.

4 Analysis: Tradeoff at the Meta-Evaluation Level

In this section, we explore the adequacy–fluency tradeoff at the meta-evaluation level. We first discuss in §4.1 how the tradeoff exists in meta-evaluation, and show that the imbalance in WMT meta-evaluation datasets imposes a bias that favors adequacy-oriented metrics. We argue that this bias consists of both an intrinsic and an extrinsic component, and that we aim to control the latter. In §4.2, we design a metric to quantify the extrinsic bias, and in §4.3, we propose a practical approach to reduce it. Finally we demonstrate in §4.4 the impact of meta-evaluation bias on the ranking of translation metrics by comparing this ranking before and after reducing the meta-evaluation extrinsic bias.

4.1 Understanding the Meta-Evaluation Bias

Translation meta-evaluation compares the ranking of translation systems produced by a given metric against the ranking based on human annotations, which in our case is the All MQM score. In this section, we will show that the All MQM ranking reflects systems' adequacy more than their fluency, leading the WMT meta-evaluation to favor adequacy-oriented metrics.

For simplicity, we assume that All MQM = Adequacy MQM+Fluency MQM.[4] Therefore, the ranking by All MQM (thus the meta-evaluation assessment) is more influenced by the component with higher variance, and we say the meta-evaluation is *biased* toward that component.

As shown in Table 2, Adequacy MQM exhibits a much higher variance than Fluency MQM in system level scores, consequently having a greater influence on the All MQM system ranking and the meta-evaluation system-level assessment. This influence is further demonstrated by the stronger alignment between All MQM and Adequacy MQM in terms of both PA and SPA. The question now is: *Should we embrace this dominance of adequacy, or is it a bias we should mitigate?*

Answering the above question requires noticing that the higher variance of the system-level Adequacy MQM scores, and consequently its greater influence, can be attributed to two possible causes:

- *Intrinsic variation,* which is due to the preference of the MQM framework and annotators. For example, adequacy errors may be generally considered more severe, leading to larger assigned penalties and thus greater variance.
- *Extrinsic variation,* which is due to the choice of translation systems during meta-evaluation. For example, we may select systems that

[3] We limit En–Zh to the conversation, e-commerce, and social domains, in line with Juraska et al. (2024).

[4] Although there are some errors that are not categorized as either adequacy or fluency, they are rare and have minor influence; therefore, we omit them for simplicity.

	Adequacy MQM				Fluency MQM				
	Variance	F-stat	PA	SPA	Variance	F-stat	PA	SPA	$B(\Delta p)$
En–De'23	0.31	36.5	0.98	0.97	0.06	7.0	0.76	0.75	0.08^A
Zh–En'23	0.23	80.6	0.97	0.93	0.03	12.9	0.70	0.74	0.03^A
En–De'24	0.24	13.7	0.88	0.87	0.11	9.5	0.85	0.81	0.04^A
En–Es'24	0.61	26.4	0.96	0.94	0.28	4.6	0.74	0.77	0.13^A
Ja–Zh'24	0.40	35.3	1.00	0.98	0.10	4.8	0.74	0.75	0.12^A
Macro-Avg	0.36	38.5	0.96	0.94	0.12	7.7	0.76	0.76	

Table 2: Adequacy MQM and Fluency MQM statistics across different evaluation sets. The reported statistics include the variance of system-level scores, the (S)PA of the scores with respect to All MQM, the F-statistic, and the B transformation of the difference between the p-values of Adequacy MQM and Fluency MQM. The last two are explained in §4.2.

differ primarily in adequacy, while performing similarly in fluency. This naturally leads to higher system-level variance in Adequacy MQM than in Fluency MQM.

Both of the above factors contribute to the variances of system-level Adequacy MQM and Fluency MQM, which in turn lead to the bias of the meta-evaluation. Therefore, we say that the meta-evaluation bias consists of two components: *intrinsic bias* and *extrinsic bias*.

The intrinsic bias reflects translation experts' beliefs about translation errors. Therefore, we retain intrinsic bias in this study. However, whether the extrinsic bias should be eliminated, retained, or partially kept is debatable and depends on the application. In our study, we aim to avoid it in order to gain a clearer understanding of the adequacy–fluency tradeoff at the evaluation level. In other contexts, one could argue that adequacy assessment should have greater influence on meta-evaluation outcomes, even at the cost of reduced fluency influence. In both cases, the extrinsic bias must first be measured separately from the intrinsic bias.

4.2 Measuring the Extrinsic Bias

To study the extrinsic bias, we propose to compare the F-statistics (Weir and Cockerham, 1984) for Adequacy MQM and Fluency MQM. In each case, the F-statistic is formulated as follows:

$$\text{F-statistic} = \frac{\text{between-system variation}}{\text{within-system variation}} \quad (3)$$

where the numerator and denominator are

$$\text{between-system variation} = \sum_{i=1}^{K} \frac{N \cdot (\bar{s}_i - \bar{s})^2}{K-1} \quad (4)$$

$$\text{within-system variation} = \sum_{i=1}^{K} \sum_{j=1}^{N} \frac{N \cdot (s_{i,j} - \bar{s}_i)^2}{N-K} \quad (5)$$

In our scenario, $s_{i,j}$ is the Adequacy MQM or Fluency MQM score of the candidate translation given by the ith system for the jth segment, $\bar{s}_i$ is the average of all the scores by the ith system, and $\bar{s}$ is the average of all the scores. N and K are the numbers of segments and systems, respectively.

The F-statistic is effective at distinguishing extrinsic variation from intrinsic variation because of how its numerator and denominator are defined. On one hand, the intrinsic variation can be regarded as a multiplicative constant within the s variables; it appears in both the numerator and denominator of Eqn. (3) and thus cancels out. On the other hand, the F-statistic captures extrinsic variation, which is due to the selected systems in the meta-evaluation. It does so by normalizing between-system variation with within-system variation.

Notice that the F-statistic depends on N and K, which makes it difficult to interpret. To this end, we adopt the ANOVA framework (Heiman, 2001) and use p-values to quantify extrinsic variation. We perform this analysis separately for Adequacy MQM and Fluency MQM.

We assume that, within each translation system, scores are independent and normally distributed around that system's mean, and that all systems share the same variance[5]; i.e., $s_{i,j} \overset{\text{iid}}{\sim} \mathcal{N}(\mu_i, \sigma^2)$. Under the null hypothesis that all systems have the same mean ($\mu_1 = \mu_2 = \cdots$), the F-statistic in Eqn. (3) follows an F distribution (Johnson et al., 1995) with degrees of freedom $K-1$ and $N-K$. The right-tail p-values are then given by

[5]ANOVA is robust to moderate variance inequality, so this assumption need only hold approximately (Lowry, 1998–2023). For completeness, we report in Appendix D the results using the variance-inequality-tolerant alternative ANOVA (Welch, 1951), which yield the same conclusions; We prefer the standard approach for its simplicity and greater numerical stability.

	Original	Synthesized by		Adequacy MQM			Fluency MQM			
		Adequacy	Fluency	F-stat	PA	SPA	F-stat	PA	SPA	$B(\overline{\Delta p})$
1	✓	–	–	38.5	0.96	0.94	7.7	0.76	0.76	0.120^{A}
2	–	✓	–	213.3	0.94	0.95	2.5	0.53	0.54	0.590^{A}
3	–	–	✓	7.5	0.62	0.71	165.1	0.82	0.78	0.450^{F}
4	✓	✓	–	111.9	0.95	0.95	4.9	0.61	0.62	0.130^{A}
5	✓	–	✓	21.7	0.80	0.83	75.5	0.73	0.73	0.030^{F}
6	–	✓	✓	93.6	0.84	0.87	72.4	0.59	0.60	$\mathbf{0.005^{A}}$
7	✓	✓	✓	72.7	0.88	0.88	48.8	0.63	0.64	$\mathbf{0.005^{A}}$

Table 3: Adequacy MQM and Fluency MQM statistics for different meta-evaluation setups, along with the B transformation of the difference between their corresponding p-values, macro-averaged across evaluation sets. Each meta-evaluation setup includes one or more system sets drawn from the original, synthesized-by-adequacy, and synthesized-by-fluency system sets.

$1 - \mathrm{CDF}(\text{F-statistic})$, where CDF denotes the cumulative distribution function of the F-distribution. Converting F-statistics to p-values puts results on a common probability scale, and makes the interpretation independent of N and K. We use this p-value as our measure of extrinsic variation.

As mentioned in §4.1, extrinsic variations of Adequacy MQM and Fluency MQM contribute to their respective variances. An asymmetry in these variations induces a bias of meta-evaluation toward adequacy or fluency, which we term *extrinsic bias*. We quantify this bias by

$$\Delta p = p_{\text{ Adequacy MQM}} - p_{\text{ Fluency MQM}} \quad (6)$$

where $p_{\text{ Adequacy MQM}}$ and $p_{\text{ Fluency MQM}}$ are the respective p-values. A positive Δp indicates bias toward adequacy, a negative Δp indicates bias toward fluency, and $\Delta p = 0$ indicates no bias between adequacy and fluency.

We observe that the range of Δp is typically very small (from 10^{-7} to 10^{-31} in Table 2), even in cases where severe bias is present (as shown in our later analysis). For presentation purposes, we report the degree of bias using $B(\Delta p) = \frac{1}{1-\log|\Delta p|}$, and append a special symbol to indicate which aspect dominates: A for adequacy ($\Delta p > 0$) and F for fluency ($\Delta p < 0$). Note that $B \in [0, 1]$, with a lower B indicating less bias.

We report the B values in Table 2, where we compare the influence of Adequacy MQM and Fluency MQM on the meta-evaluation in five evaluation datasets. Results show a consistent extrinsic bias toward adequacy. We argue that this behavior stems from the particular composition of translation systems in the WMT datasets, and may be controlled by carefully selecting or synthesizing systems with more variance in their Fluency MQM.

4.3 Reducing the Extrinsic Bias

In this subsection, we provide a method to reduce meta-evaluation extrinsic bias. This not only has practical values, but also allows us to better evaluate the adequacy–fluency bias of existing translation evaluation metrics (to be discussed in §5).

We accomplish this by considering additional pseudo-translation systems in the meta-evaluation. In particular, we synthesize two sets of translation systems: one exhibiting extreme variations in system-level Adequacy MQM, and the other in Fluency MQM. This design helps to control the variation in each dimension and thereby reduces extrinsic bias through controlled mixing of candidate translation systems.

Suppose we have K original translation systems. We synthesize K adequacy-oriented systems in the following way: for each segment, we assume that the kth synthesized system generates the kth most adequate translation (according to Adequacy MQM), among the original K translation systems' outputs.[6] It is easy to see that, given the available translation candidates, such K synthesized translate systems exhibit extreme variation in Adequacy MQM at the system level. Likewise, we synthesize K fluency-oriented systems, and in total, we have $3K$ translation systems for meta-evaluation. It is emphasized that our research does not require additional annotation effort, as the above synthesizing process utilizes readily available MQM scores.

Table 3 illustrates how synthesized systems help control the extrinsic bias. The table reports the macro-average of metrics across the five evaluation sets (§3.1).

[6]For segments with tied scores, we rank them randomly.

Rows 2 and 3 represent extreme configurations that maximize the F-statistic for Adequacy MQM and Fluency MQM, respectively. We see that they yield extreme B values, with Row 2 being biased toward adequacy and Row 3 toward fluency. This expected pattern confirms that (1) the B value effectively captures extrinsic bias arising from system selection, and (2) our synthesis method can control such a bias.

We observe that the setup in Row 5 has an overall bias toward adequacy (indicated by higher PA and SPA scores), although the extrinsic bias is toward fluency (indicated by the F annotation). This suggests the existence of intrinsic bias discussed in §4.1, further justifying the need of separately quantifying the intrinsic and extrinsic bias.

We also observe that Rows 6 and 7 yield the lowest B values, and we consider them the most balanced meta-evaluation setups, which will be used in our study of adequacy–fluency tradeoff at the evaluation level in §5.

4.4 The Effect of Meta-Evaluation Bias on Metric Comparisons

In the previous parts, we have demonstrated that the standard WMT meta-evaluation is biased toward adequacy, and we propose to carefully synthesize pseudo-translation systems to control this bias. In this part, we show how the meta-evaluation bias can influence the development of translation metrics.

Table 4 presents how the original WMT setup (Row 1 of Table 3) and *our balanced setup* (Row 6 of Table 3) rank various metrics, using PA and SPA as the meta-metrics.

It is interesting to examine "CometKiwi 22 XXL" and "MetricX (ours)" in detail, shown by the bold lines in Table 4. As seen, CometKiwi 22 XXL consistently and significantly outperforms MetricX (ours) under the original meta-evaluation, which is biased toward adequacy; however, this trend is reversed under our balanced meta-evaluation setup. In §5, we will show that CometKiwi 22 XXL has a considerable bias toward adequacy, whereas MetricX (ours) demonstrates a more balanced behavior. This comparison shows that a metric is favored if its bias in the adequacy–fluency spectrum aligns with that of the meta-evaluation; consequently, the development of translation metrics inherits the bias of the meta-evaluation. Our analysis highlights the importance of studying translation meta-evaluation imbalance.

Metric	Original setup PA	Original setup SPA	Our balanced setup PA	Our balanced setup SPA
AdequacyX	0.881 4	0.866 4	0.847 1	0.833 1
AdequacyX QE	0.885 3	0.873 1	0.834 2	0.824 5
FluencyX	0.697 13	0.720 12	0.676 12	0.675 13
MetricX (ours)	**0.862** 7	**0.845** 8	**0.833** 3	**0.829** 3
MetricX QE (ours)	0.878 6	0.861 6	0.809 6	0.823 6
MetricX-24	0.880 5	0.866 3	0.820 5	0.831 2
MetricX-24 QE	0.888 2	0.867 2	0.820 4	0.827 4
Comet 22	0.850 8	0.847 7	0.800 8	0.804 8
CometKiwi 22	0.813 9	0.802 9	0.782 9	0.772 9
CometKiwi 22 XXL	**0.889** 1	**0.862** 5	**0.805** 7	**0.815** 7
Gemma 3 (4B)	0.701 12	0.754 10	0.652 13	0.736 10
BLEU (sent. level)	0.726 11	0.712 13	0.726 10	0.689 12
ChrF (sent. level)	0.739 10	0.731 11	0.722 11	0.721 11

Table 4: Meta-evaluation of translation metrics using the original WMT setup and our balanced setup. Decimal numbers denote the PA and SPA scores, macro-averaged across our five evaluation sets; blue squares indicate the corresponding rankings. We bold MetricX (ours) and CometKiwi 22 XXL, as they are discussed in detail in the main text. Note that the results here are not identical to those in the WMT reports as we are reporting the average results across specific evaluation sets.

Evaluation Set	Concordance	Discordance
En–De'23	136 (49%)	140 (51%)
Zh–En'23	124 (29%)	311 (71%)
En–De'24	282 (50%)	279 (50%)
En–Es'24	139 (43%)	186 (57%)
Ja–Zh'24	156 (51%)	149 (49%)

Table 5: Concordance and discordance between adequacy and fluency in our balanced setup, reported as system pair counts and percentages.

5 Analysis: Tradeoff at the Evaluation Level

In this section, we analyze the adequacy–fluency tradeoff at the evaluation level and determine the bias of each metric within this tradeoff. Note that the notion of bias considered here is related to but distinct from that in §4. We say a metric is biased toward adequacy or fluency if it ranks translation systems in a way that is more aligned with the ranking derived solely from that aspect, as measured by the corresponding MQM scores.

To study the aforementioned bias, we would like to separately meta-evaluate the adequacy and fluency of each metric, using the original WMT meta-evaluation setup and our balanced setup described in Row 6 of Table 3.[7] Notice that the latter is a

[7]Although Rows 6 and 7 in Table 3 have similar B values, we prefer the former because it is fully synthesized and has a

Metric	Original setup			Our balanced setup		
	Agreement in discordant cases		PA in concordant cases	Agreement in discordant cases		PA in concordant cases
	with Adequacy MQM	with Fluency MQM		with Adequacy MQM	with Fluency MQM	
All MQM	86%	14%	100%	74%	26%	100%
Adequacy MQM	100%		100%	100%		100%
Fluency MQM		100%	100%		100%	100%
AdequacyX	65%	35%	93%	70%	30%	91%
AdequacyX QE	70%	30%	93%	69%	31%	91%
FluencyX	22%	78%	84%	44%	56%	77%
MetricX (ours)	58%	42%	93%	69%	31%	90%
MetricX QE (ours)	61%	39%	95%	66%	34%	89%
MetricX-24	64%	36%	94%	70%	30%	89%
MetricX-24 QE	66%	34%	94%	70%	30%	89%
Comet 22	67%	33%	90%	73%	27%	86%
CometKiwi 22	72%	28%	84%	76%	24%	81%
CometKiwi 22 XXL	78%	22%	92%	73%	27%	87%
Gemma 3 (4B)	38%	62%	81%	47%	53%	76%
BLEU (sentence level)	71%	29%	74%	59%	41%	79%
ChrF (sentence level)	71%	29%	76%	65%	35%	77%

Table 6: PA breakdown for the original setup and our balanced setup, macro-averaged across five evaluation sets. For each metric, the agreements with Adequacy MQM and Fluency MQM sum to 1, by formulation and because ties between system-level Adequacy MQM and Fluency MQM scores are rare.

synthetic setup, which may not be representative of real-world situations. It is included only for our adequacy–fluency analysis.

From the data statistics in Table 5, we observe an even stronger disagreement between adequacy- and fluency-based system rankings under the balanced setup than under the original setup (Table 1). This is expected, as the balanced setup synthesizes the most- and least-adequate systems, which are unlikely to correspond to the most- and least-fluent systems, and vice versa.

Despite the significant discordance between Adequacy MQM and Fluency MQM, we also observe a large number of concordant cases (i.e., one system outperforms another in both Adequacy MQM and Fluency MQM). This raises a challenge when we separately analyze adequacy and fluency. To address this, we design three experimental protocols, as follows.

5.1 PA Breakdown

In this part, we design an experimental protocol using PA to separately analyze how a metric assesses adequacy and fluency, without being misled by the concordant cases. Since PA counts binary agreements, we may simply exclude concordant cases from our adequacy/fluency analysis.

Specifically, we first split our system pairs into two disjoint subsets based on whether their Adequacy MQM and Fluency MQM are concordant or discordant. Then, for each metric we report:

- PA in concordant cases, measuring the metric's general performance;
- PA with respect to Adequacy MQM in discordant cases, which measures the metric's bias toward adequacy (thus we call it "agreement with Adequacy MQM"); and
- PA with respect to Fluency MQM in discordant cases, which is called "agreement with Fluency MQM."

In Table 6, we present the results macro-averaged[8] across different evaluation sets.[9]

As seen, both BLEU and ChrF achieve more agreement with Adequacy MQM than Fluency MQM. This provides empirical evidence for the belief in the literature that BLEU and ChrF lean toward adequacy (Flamich et al., 2025). On the other hand, FluencyX and Gemma 3 exhibit a stronger bias toward fluency, which aligns with their design. Moreover, All MQM shows a strong alignment with Adequacy MQM in the original setup and a weaker alignment in our balanced setup, confirming our claims in §4 that current WMT meta-evaluation is biased toward adequacy and that such bias is partially mitigated in our balance meta-evaluation setup.

clearer construction. For completeness, we also report results for the setup in Row 7 in Appendix E.

[8] In our preliminary experiments, we also analyzed the micro-averaged results, which led to the same conclusions as those obtained here.

[9] We report results for individual evaluation sets under the original setup in Appendix F, where we observe diverse behaviors by the metrics. This variation may be due to the noise of metric performance, which tends to be smoothed out when averaged across evaluation sets.

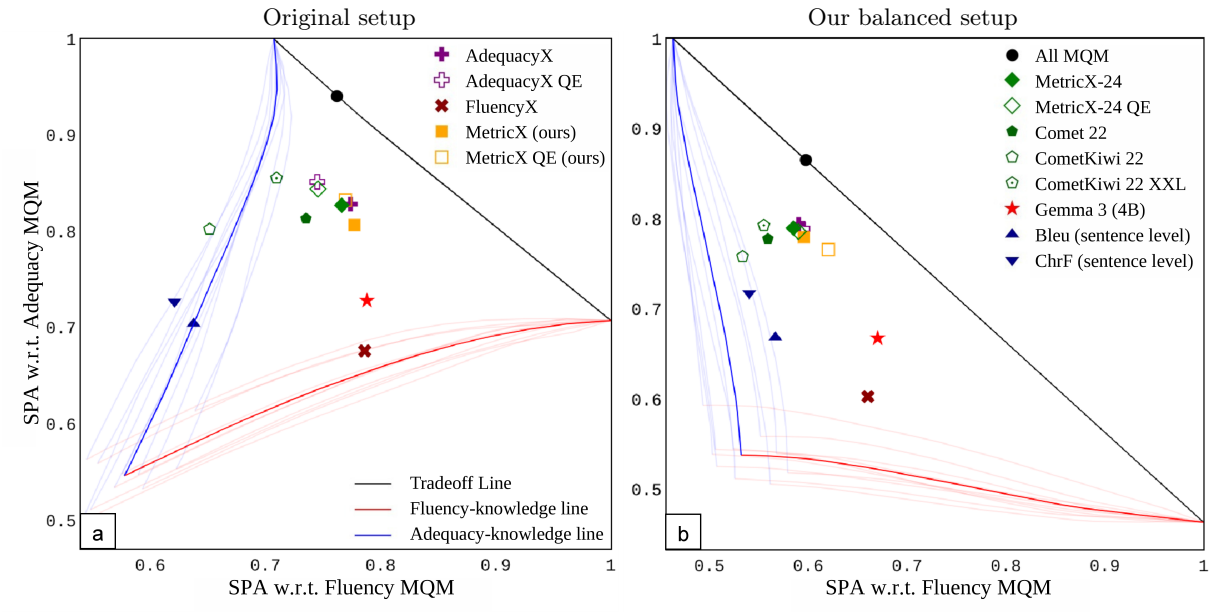

Figure 1: SPA plane using (a) the original setup and (b) our balanced setup. Legends are shared between the two figures. Each shadow line represents the combination of Adequacy MQM or Fluency MQM with a specific random noise instance. The solid red and blue lines are the average of the shadow lines.

Interestingly, FluencyX and Gemma 3 show some alignment with adequacy, despite lacking access to the source texts. This can be explained by two possibilities: (1) These metrics are imperfect and fail to align entirely with Fluency MQM, although designed to do so. (2) When the metrics are trained, they have learned strong prior for adequacy errors (e.g., due to the adequacy–fluency correlation of training segments). This aligns with the findings of Perrella et al. (2024), who show that metrics based solely on candidate translations or source texts can achieve high performance.

For most other metrics, we find that they exhibit stronger alignment with Adequacy MQM than with Fluency MQM. Among them, the MetricX variants display a relatively more balanced behavior than other metrics, including the Comet variants.

Finally, we would like to point out that our analysis sheds light on the position of different evaluation metrics within the adequacy–fluency tradeoff. However, whether a metric should achieve 50%–50% balance or mimicking All MQM (which is highly biased toward adequacy but often treated as the evaluation ground truth) is a task-specific design choice.

5.2 SPA Plane

We also aim to conduct an adequacy–fluency analysis using SPA. However, we cannot mirror the PA breakdown in §5.1, because SPA lacks a binary notion of concordance versus discordance. Instead, we perform qualitative visualization to demonstrate the adequacy–fluency tradeoff of a metric.

Specifically, we design a plot where the x-axis and y-axis represent SPA computed with respect to Fluency MQM and Adequacy MQM, respectively. Each metric is shown as a single point in this space. We then augment the plot with three sentinel lines:

- *Tradeoff line* (the black lines in Figure 1), representing the linear interpolation of Adequacy MQM and Fluency MQM. No system can surpass the tradeoff line.
- *Adequacy-knowledge line* (the blue lines in Figure 1), derived from linear interpolation of Adequacy MQM and random scores (i.e., uniform in the range of Adequacy MQM scores for each segment). The interpolation is accomplished by weighted average of the scores. Points on this line achieve fluency scores solely due to the correlation between Adequacy MQM and Fluency MQM.
- *Fluency-knowledge line* (the red lines in Figure 1), derived from linear interpolation of Fluency MQM and the random score.

The latter two lines are averaged from 10 computations with different random score instances.

These three lines form a triangle-like shape with vertices at Adequacy MQM, Fluency MQM, and pure random noise. A metric's closeness to the tradeoff line indicates its general quality, while its closeness to the other two lines reveals its bias toward adequacy or fluency.

Metric	Unnormalized sensitivity		Normalized sensitivity	
	Adequacy	Fluency	Adequacy	Fluency
All MQM	0.995	0.998	0.882	0.353
Adequacy MQM	1.000	0.000	1.000	0.000
Fluency MQM	0.000	1.000	0.000	1.000
AdequacyX	0.211	0.146	0.667	0.184
AdequacyX QE	0.143	0.084	0.474	0.111
FluencyX	0.010	0.018	0.202	0.145
MetricX (ours)	0.125	0.100	0.644	0.206
MetricX QE (ours)	0.081	0.070	0.521	0.180
MetricX-24	0.379	0.257	0.591	0.160
MetricX-24 QE	0.300	0.207	0.503	0.139
Comet 22	0.010	0.006	0.593	0.142
CometKiwi 22	0.007	0.003	0.415	0.071
CometKiwi 22 XXL	0.017	0.010	0.489	0.122
Gemma 3 (4B)	0.037	0.101	0.212	0.231
BLEU (sent. level)	1.201	0.755	0.383	0.099
ChrF (sent. level)	1.088	0.657	0.149	0.090

Table 7: Normalized and unnormalized sensitivity against adequacy and fluency.

Figure 1 presents the results, also macro-averaged across the evaluation sets in §5.1. The black line illustrates the severity of the adequacy–fluency tradeoff in translation evaluation (given a certain meta-evaluation setup): a larger top-right area indicates a more severe tradeoff, as this area is not reachable. As shown, our balanced setup exhibits a more severe tradeoff, which is consistent with Table 5. Moreover, our balanced setup has a larger angle between the blue and red lines, indicating a lower correlation between the two aspects.

Regarding specific metrics, we observe that FluencyX and Gemma 3 lie close to the fluency boundary (red line), while other metrics (namely, BLEU, ChrF, Comet, and MetricX) are biased toward adequacy. In particular, Comet variants are more biased than MetricX variants. The SPA results are consistent with the analysis using PA (§5.1).

5.3 Sensitivity Analysis

In §5.1 and §5.2, we illustrate the position of each metric in the adequacy–fluency tradeoff by their system-level ranking prediction. Here, we aim to analyze how each metric reacts to an adequacy or fluency error.

We do this by controlling one aspect while varying the other. Take adeqacy as an example. Given a source segment, we consider pairs of translation candidates that share the same Fluency MQM but differ in Adequacy MQM. For a metric, we report $\frac{\Delta \text{ metric score}}{\Delta \text{ Adequacy MQM}}$, averaged over all our translation pairs. This measures the expected change in the metric score per one-point difference in Adequacy MQM.

We also report a normalized version of the above measure by multiplying it with $\frac{\sum \sigma(\text{Adequacy MQM})}{\sum \sigma(\text{metric score})}$, where $\sigma(\cdot)$ is the standard deviation over different candidate translations given a segment, and the sum is taken over all segments. This scaling enables meaningful comparison across metrics with different output scales.

Table 7 reports the results for both adequacy and fluency. Here, we can interpret the bias of a metric by comparing the sensitivity to adequacy and that to fluency. Generally, our findings in previous subsections are also observed in this experiment, except for the normalized sensitivity of FluencyX, which requires further investigation.

Overall, this section provides three analysis protocols (PA Breakdown, SPA Plain, and Sensitivity Analysis) to study the position of different evaluation metrics within the adequacy–fluency tradeoff. Our key findings in this section include: (1) Most of the translation metrics are biased toward adequacy, and (2) For the commonly used MetricX and Comet metrics, the former exhibits a more balanced behavior than the latter. Consistent observations in different protocols cross-validate the reliability of our findings.

6 Conclusion

In this work, we investigate the adequacy–fluency tradeoff in translation. While this tension is well-documented at the level of translation output (Flamich et al., 2025), we show that it also manifests severely at the levels of evaluation and meta-evaluation.

Through empirical analysis of WMT meta-evaluation protocols, we uncover a systematic bias toward adequacy, driven by the composition of meta-evaluation datasets. We also propose a method that can control the balance between adequacy and fluency.

We further analyze the placement of popular translation evaluation metrics along the adequacy–fluency tradeoff and find that most metrics lean toward adequacy.

We argue that the adequacy–fluency tradeoff is a critical yet under-recognized matter in translation evaluation and meta-evaluation. While we do not take a stance on how to deal with such bias, our primary contribution is to raise awareness of this matter within the community.

We discuss future work in Appendix G.

Limitations

Our work provides an extensive investigation of adequacy–fluency trade-offs in the evaluation and meta-evaluation of machine translation. However, it also has limitations.

First, our work is based on MQM data, which includes human evaluation that is inherently noisy. It places limits on the generality of the conclusion we draw in this paper. To mitigate this, we report results macro-averaged over five translation datasets. The results on individual dataset are considerably noisier, as shown in Appendix F.

Second, this study covers only a limited set of language pairs and systems, owing to the limited availability of MQM data. Expanding the evaluation data would yield more robust conclusion.

Third, we only analyze extrinsic bias (caused by the composition of translation systems in meta-evaluation), while not addressing the intrinsic bias (caused by the design of the MQM framework). Studying the intrinsic bias is highly impactful and warrants further efforts.

References

Sweta Agrawal, George Foster, Markus Freitag, and Colin Cherry. 2021. Assessing reference-free peer evaluation for machine translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1158–1171.

Bogdan Babych and Anthony Hartley. 2008. Sensitivity of automated MT evaluation metrics on higher quality MT output: BLEU vs task-based evaluation methods. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 2133–2136.

Rafael E. Banchs, Luis F. D'Haro, and Haizhou Li. 2015. Adequacy–fluency metrics: Evaluating MT in the continuous space model framework. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(3):472–482.

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.

Eirini Chatzikoumi. 2020. How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering*, 26(2):137–161.

Shweta Chauhan and Philemon Daniel. 2023. A comprehensive survey on various fully automatic machine translation evaluation metrics. *Neural Processing Letters*, page 12663–12717.

Daniel Deutsch, George Foster, and Markus Freitag. 2023. Ties matter: Meta-evaluating modern metrics with pairwise accuracy and tie calibration. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 12914–12929.

Gergely Flamich, David Vilar, Jan-Thorsten Peter, and Markus Freitag. 2025. You cannot feed two birds with one score: The accuracy–naturalness tradeoff in translation. *arXiv preprint arXiv:2503.24013*.

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.

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*, pages 46–68.

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.

Gemma Team. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

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.

Gary W Heiman. 2001. *Understanding Research Methods and Statistics: An Integrated Introduction for Psychology*. Houghton, Mifflin and Company.

Takumi Ito, Kees van Deemter, and Jun Suzuki. 2025. Reference-free evaluation metrics for text generation: A survey. *arXiv preprint arXiv:2501.12011*.

Norman L Johnson, Samuel Kotz, and Narayanaswamy Balakrishnan. 1995. *Continuous Univariate Distributions*, volume 2. John Wiley & Sons.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 Metrics Shared Task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 Metrics Shared Task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494.

Alon Lavie and Michael J Denkowski. 2009. The METEOR metric for automatic evaluation of machine translation. *Machine Translation*, 23(2):105–115.

Richard Lowry. 1998–2023. *Concepts & Applications of Inferential Statistics*. Online Book.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 Metrics Shared Task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers)*, pages 62–90.

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.

Valerie Mariana, Troy Cox, and Alan Melby. 2015. The multidimensional quality metrics (MQM) framework: A new framework for translation quality assessment. *The Journal of Specialised Translation*, pages 137–161.

Marianna Martindale and Marine Carpuat. 2018. Fluency over adequacy: A pilot study in measuring user trust in imperfect MT. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 13–25.

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.

Nikita Moghe, Arnisa Fazla, Chantal Amrhein, Tom Kocmi, Mark Steedman, Alexandra Birch, Rico Sennrich, and Liane Guillou. 2025. Machine translation meta evaluation through translation accuracy challenge sets. *Computational Linguistics*, 51(1):73–137.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Edoardo Barba, and Roberto Navigli. 2024. Guardians of the machine translation meta-evaluation: Sentinel metrics fall in! In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 16216–16244.

John R. Pierce and John B. Carroll. 1966. *Language and Machines: Computers in Translation and Linguistics*. National Academy of Sciences/National Research Council, USA.

Maja Popović. 2015. chrF: Character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 578–585.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848.

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*, pages 2685–2702.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. CometKiwi: IST-Unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 634–645.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. RUBER: An unsupervised method for automatic evaluation of open-domain dialog systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 722–729.

Brian Thompson, Nitika Mathur, Daniel Deutsch, and Huda Khayrallah. 2024. Improving statistical significance in human evaluation of automatic metrics via soft pairwise accuracy. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1222–1234.

B. S. Weir and C. Clark Cockerham. 1984. Estimating f-statistics for the analysis of population structure. *Evolution*, 38(6):1358–1370.

B. L. Welch. 1951. On the comparison of several mean values: An alternative approach. *Biometrika*, 38(3/4):330–336.

William J. Welch. 1990. Construction of permutation tests. *Journal of the American Statistical Association*, 85(411):693–698.

John S. White and Theresa A. O'Connell. 1993. Evaluation of machine translation. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. 2019. Beyond BLEU: Training neural machine translation with semantic similarity. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 4344–4355.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

A Adequacy MQM and Fluency MQM Categorization

As mentioned in §3.1, we consider Adequacy MQM and Fluency MQM separately, following the categorizations provided in Flamich et al. (2025). Table 8 shows the categorization used for En–De, Ja–Zh, and Zh–En. Specially, En–Es follows the categorization in Table 9.

B Dataset Statistics

Table 10 summarizes key statistics for each evaluation dataset in our study.

C PA vs. SPA

In the main paper, we mention that PA and SPA could behave differently, and that it is necessary to consider both PA and SPA in our study. In this appendix, we use a toy example to illustrate the potential disagreement of PA and SPA when assessing whether a metric is biased toward Adequacy MQM or Fluency MQM.

Table 11 shows the details of our example. In this case, PA suggests that the metric is biased toward adequacy, because PA only judges based on the sign. However, SPA suggests the opposite, as it concerns the closeness of the score difference. See §2.2 for details.

Therefore, we use both of the metrics, and design dedicated experiments for them. Luckily, our findings are generally consistent under PA and SPA, suggesting that they are reliable meta-metrics in our study.

D Measuring Extrinsic Bias without Equal-Variance Assumption

In §4.2, for simplicity we assume that all translation systems share the same variance in their Adequacy MQM scores and the same variance in their Fluency MQM scores. In this part, we drop this assumption and use Welch (1951)'s ANOVA approach (which does not require the equal-variance assumption) to calculate F-statistics and B values.

Table 12 reports our results. The findings are consistent with those reported in Table 2 under the equal-variance assumption. We nevertheless prefer the standard approach presented in the main text due to its greater simplicity and numerical stability. Note that we observe undefined p-values in our preliminary experiments for some synthetic datasets in §4.3, if we do not have the assumption.

E Analysis Results for Setup 7

In this section, we present the results of PA Breakdown (§5.1) and SPA Plane (§5.2) based on Setup 7 in Table 3. We include this setup for the sake of completeness, as it closely matches Row 6 (used in our main analysis) in terms of B values. Table 13, Table 14, and Figure 2 present the concordance/discordance ratio, PA Breakdown, and SPA Plane results, respectively, for Setup 7. All our findings based on Setup 6 also hold in this setup.

F Adequacy–Fluency Tradeoff Results on Individual Evaluation Sets

In §5, we report results macro-averaged across five evaluation sets (of different language pairs). In this part, we present the results on each evaluation set.

In Table 15 and Figure 3, we report PA-Breakdown and SPA Planes results based on the original WMT setup, presented separately for each evaluation set. The results exhibit diverse behaviors across evaluation sets. Thus, we perform macro-average in our main paper. The noisy phenomenon requires further investigation.

G Future Work

Our study opens several avenues for future research, discussed below.

Accuracy errors	Fluency errors		Other
Accuracy/Addition	Fluency/Grammar	Style/Archaic or obscure word choice	Other
Accuracy/Creative Reinterpretation	Fluency/Inconsistency	Style/Bad sentence structure	Source issue
Accuracy/Gender Mismatch	Fluency/Punctuation	Style/Unnatural or awkward	
Accuracy/Mistranslation	Fluency/Register	Locale convention/Address format	
Accuracy/Omission	Fluency/Spelling	Locale convention/Currency format	
Accuracy/Source language fragment	Fluency/Text-Breaking	Locale convention/Time format	
Non-translation!	Terminology/Inconsistent	Terminology/Inappropriate for context	

Table 8: Flamich et al. (2025)'s categorization of MQM errors, used for En–De, Ja–Zh, and Zh–En translation directions.

Accuracy errors	Fluency errors		Other
Addition	Capitalization	Date-time format	Other
Agreement	Inconsistency	Lacks creativity	Source issue
Do not translate	Grammar	Measurement format	
Mistranslation	Number format	Punctuation	
MT hallucination	Register	Spelling	
Omission	Unnatural flow	Whitespace	
Untranslated	Word order	Wrong language variety	
Wrong named entity			
Wrong term			

Table 9: Flamich et al. (2025)'s categorization of MQM errors, used for the En–Es translation direction.

Year	2023	2023	2024	2024	2024
Language pair	En–De	Zh–En	En–De	En–Es	Ja–Zh
Mean All MQM	6.17	2.51	2.50	0.58	2.92
Mean Adequacy MQM	3.97	1.62	1.38	0.45	2.55
Mean Fluency MQM	1.99	0.89	1.10	0.13	0.35
Mean non-zero All MQM	10.02	5.20	4.95	3.00	6.00
Mean non-zero Adequacy MQM	9.39	6.21	5.63	3.91	6.38
Mean non-zero Fluency MQM	4.23	2.36	2.82	1.37	2.07
# segments	5520	1954	8766	8722	7840
# segments w/o errors (All MQM = 0)	1350	350	3901	6672	3842
# segments w/ errors (All MQM > 0)	4170	1604	4865	2050	3998
# segments w/ adequacy errors (Adequacy MQM > 0)	2919	891	2738	1393	3354
# segments w/ fluency errors (Fluency MQM > 0)	3149	1250	3462	849	1325

Table 10: Dataset statistics by language pair.

First, we reveal the potential imbalance in the translation meta-evaluation and highlight the importance of understanding this phenomenon (§4). We propose a statistical measure to quantify this imbalance and design a method to control it through data synthesis. However, there is room for improvement in both measuring the imbalance and debiasing the meta-evaluation. In particular, we are interested in exploring theoretical approaches to debiasing the meta-evaluation outputs by normalizing scores in a post hoc manner, without altering the datasets.

Second, we develop dedicated translation metrics, AdequacyX and FluencyX, to independently assess adequacy and fluency (§3.3). We aim to further improve this separation, as it facilitates adequacy–fluency analyses, especially in scenarios where MQM annotations are not available.

Third, as we are now in the era of large language models (LLMs), it is interesting to study the LLM-as-a-judge for translation through the lens of the adequacy–fluency tradeoff, for example, understanding and steering the bias of LLM-as-a-judge.

	Sys 1	Sys 2	Δ
Adequacy MQM	8.0	6.0	+2.0
Fluency MQM	6.0	6.5	−0.5
Metric	6.3	6.0	+0.3

closer

same sign

Table 11: A toy example illustrating potential disagreement between PA and SPA: PA prefers keeping the same sign, whereas SPA prefers the closer one. Here, all numbers are hypothetical for illustration purposes.

	F-statistic		$B(\Delta p)$
	Adequacy	Fluency	
En–De'23	28.9	8.4	0.07^A
Zh–En'23	84.1	11.9	0.04^A
En–De'24	13.0	9.0	0.04^A
En–Es'24	26.5	4.2	0.14^A
Ja–Zh'24	28.8	6.9	0.08^A
Macro-Avg	36.3	8.1	

Table 12: Welch (1951)'s F-statistics for Adequacy MQM and Fluency MQM, and the respective B-values, in the original meta-evaluation setup.

Evaluation Set	Concordance	Discordance
En–De'23	356 (57%)	274 (43%)
Zh–En'23	370 (37%)	620 (63%)
En–De'24	692 (54%)	583 (46%)
En–Es'24	375 (51%)	366 (49%)
Ja–Zh'24	399 (54%)	342 (46%)

Table 13: Concordance and discordance between Adequacy MQM and Fluency MQM in Setup 7, reported as system pair counts and percentages.

	Setup 7	
Metric	Agreement in discordant cases with Adequacy MQM / with Fluency MQM	PA in concordant cases
All MQM	76% / 24%	100%
Adequacy MQM	100%	100%
Fluency MQM	100%	100%
AdequacyX	67% / 33%	92%
AdequacyX QE	67% / 33%	92%
FluencyX	41% / 59%	79%
MetricX (ours)	66% / 34%	92%
MetricX QE (ours)	64% / 36%	91%
MetricX-24	67% / 33%	91%
MetricX-24 QE	67% / 33%	91%
Comet 22	69% / 31%	88%
CometKiwi 22	72% / 28%	83%
CometKiwi 22 XXL	71% / 29%	90%
Gemma 3 (4B)	46% / 54%	78%
BLEU (sentence level)	61% / 39%	78%
ChrF (sentence level)	64% / 36%	78%

Table 14: PA breakdown, macro-averaged across five evaluation sets, for Setup 7.

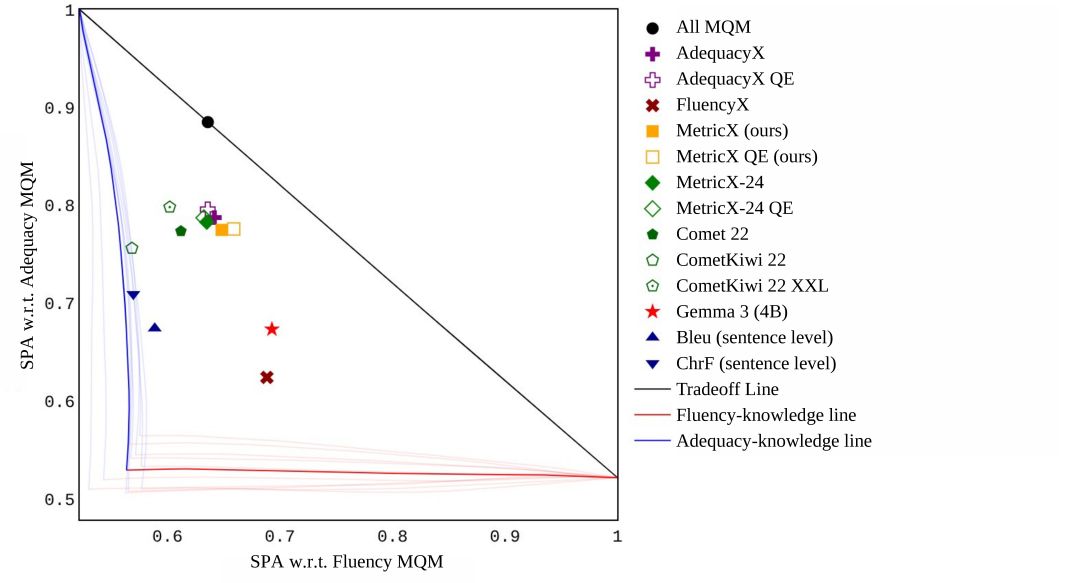

Figure 2: SPA plane using Setup 7, macro-averaged across five evaluation sets. Each shadow line represents the combination of Adequacy MQM or Fluency MQM with a specific random noise instance. The solid red and blue lines are the average of the shadow lines.

Metric	En-De 2023			Zh-En 2023		
	Agreement in discordant cases		PA in concordant cases	Agreement in discordant cases		PA in concordant cases
	with Adequacy MQM	with Fluency MQM		with Adequacy MQM	with Fluency MQM	
All MQM	94%		100%	91%		100%
Adequacy MQM	100%		100%	100%		100%
Fluency MQM		100%	100%		100%	100%
AdequacyX	76%	24%	100%	77%	23%	94%
AdequacyX QE	88%		100%	71%	29%	96%
FluencyX	35%	65%	92%		91%	89%
MetricX (ours)	65%	35%	100%	66%	34%	87%
MetricX QE (ours)	71%	29%	100%	66%	34%	96%
MetricX-24	94%		98%	74%	26%	96%
MetricX-24 QE	94%		98%	71%	29%	96%
Comet 22	88%		98%	80%	20%	86%
CometKiwi 22	94%		96%	71%	29%	93%
CometKiwi 22 XXL	94%		98%	74%	26%	94%
Gemma 3 (4B)	65%	35%	90%		91%	73%
BLEU (sentence level)	88%		88%	80%	20%	77%
ChrF (sentence level)	88%		82%	71%	29%	80%

Metric	En-De 2024			En-Es 2024		
	Agreement in discordant cases		PA in concordant cases	Agreement in discordant cases		PA in concordant cases
	with Adequacy MQM	with Fluency MQM		with Adequacy MQM	with Fluency MQM	
All MQM	55%	45%	100%	87%	13%	100%
Adequacy MQM	100%		100%	100%		100%
Fluency MQM		100%	100%		100%	100%
AdequacyX	42%	58%	92%	57%	43%	87%
AdequacyX QE	53%	47%	90%	70%	30%	87%
FluencyX	16%	84%	94%	26%	74%	71%
MetricX (ours)	45%	55%	94%	57%	43%	89%
MetricX QE (ours)	50%	50%	95%	65%	35%	89%
MetricX-24	42%	58%	92%	48%	52%	85%
MetricX-24 QE	53%	47%	97%	52%	48%	85%
Comet 22	58%	42%	95%	48%	52%	84%
CometKiwi 22	82%	18%	78%	57%	43%	69%
CometKiwi 22 XXL	74%	26%	89%	65%	35%	87%
Gemma 3 (4B)	47%	53%	82%	48%	52%	73%
BLEU (sentence level)	63%	37%	80%	43%	57%	53%
ChrF (sentence level)	71%	29%	80%	43%	57%	64%

Metric	Ja-Zh 2024		
	Agreement in discordant cases		PA in concordant cases
	with Adequacy MQM	with Fluency MQM	
All MQM	100%		100%
Adequacy MQM	100%		100%
Fluency MQM		100%	100%
AdequacyX	75%	25%	93%
AdequacyX QE	70%	30%	91%
FluencyX	25%	75%	72%
MetricX (ours)	60%	40%	97%
MetricX QE (ours)	55%	45%	95%
MetricX-24	60%	40%	97%
MetricX-24 QE	60%	40%	95%
Comet 22	60%	40%	88%
CometKiwi 22	55%	45%	86%
CometKiwi 22 XXL	85%	15%	90%
Gemma 3 (4B)	20%	80%	86%
BLEU (sentence level)	80%	20%	72%
ChrF (sentence level)	80%	20%	76%

Table 15: PA breakdown, per evaluation set, based on the original WMT setup.

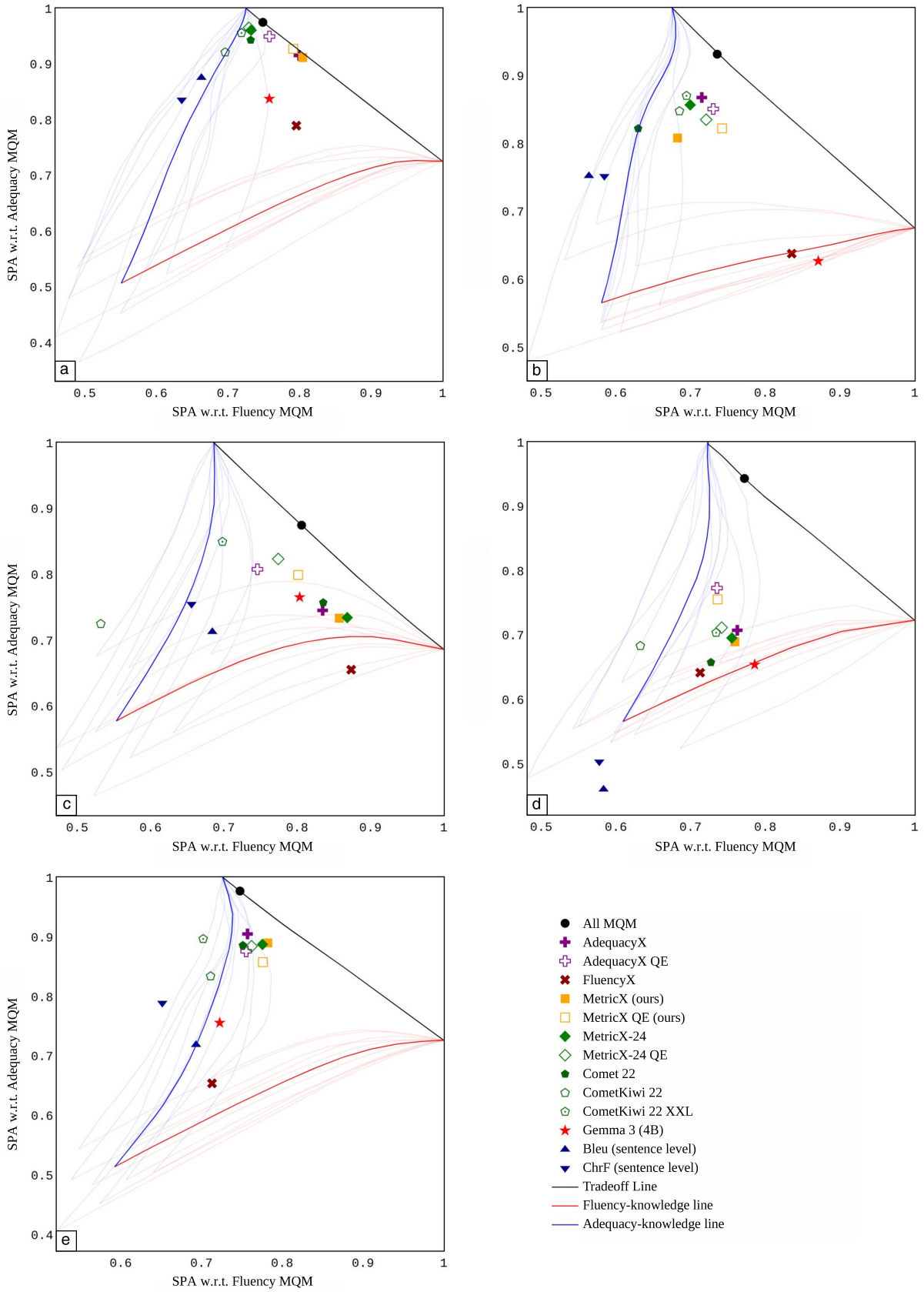

Figure 3: SPA plane based on the original WMT setup for (a) En–De 2023, (b) Zh–En 2023, (c) En–De 2024, (d) En–Es 2024, and (e) Ja–Zh 2024. The legend is shared for all the figures.