

Mind Which Bird You Favour: Parameterizing Adequacy–Fluency Balance in Meta-Evaluation of Machine Translation

Behzad Shayegh and Niloofar Kazemi

Independent Researchers

Abstract

There is a tradeoff in machine translation meta-evaluation between prioritizing alignment with adequacy versus fluency. The balance depends on the combination of translation systems in the meta-evaluation dataset. This system set is a small, filtered sample whose characteristics change heavily across years and language pairs; it does not represent the true system distribution. Consequently, the adequacy–fluency balance is often unrepresentative and subject to change. For sensitive domains, controlling this balance is critical. We expose this balance as a tunable choice. To achieve a target balance, we reweight existing systems while minimizing distortion from uniform weighting, ensuring the evaluated systems remain real and representative. We provide an exact optimization algorithm with theoretical guarantees and pruning mechanisms to compute these weights. To validate meta-evaluation internal consistency, we design a scorer-augmentation framework that establishes a known relative identity for the scorers. Results demonstrate that our reweighting method effectively controls the adequacy–fluency balance and preserves the internal consistency, outperforming prior approaches. Finally, we analyze the performance of popular scorers across a sweep of this parameter.¹

1 Introduction

Machine translation (MT) is typically judged by automatic metrics, which we refer to as *scorers*.² Deciding which scorer to trust is itself an evaluation problem known as *meta-evaluation* (Callison-Burch et al., 2007). Concretely, meta-evaluation fixes a set of translation systems and scores them once with a human protocol taken as the ground

*Correspondence to: the.shayegh@gmail.com

¹Code available at <https://github.com/TheShayegh/adequacy-fluency-param>

²We use the term “scorer” instead of “metric” to avoid ambiguity with meta-metrics.

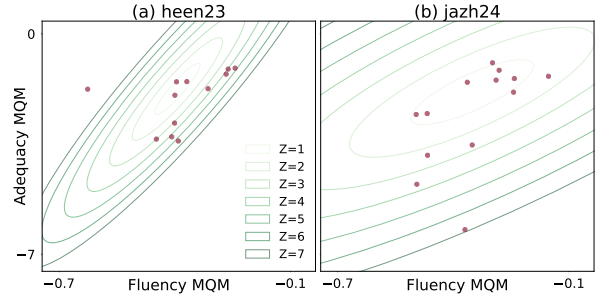


Figure 1: System-level Adequacy vs. Fluency MQM for (a) heen23 and (b) jazh24 WMT datasets. Each dot is an MT system. Green contours are robust Mahalanobis-distance levels (Gnanadesikan and Kettenring, 1972).

truth (Multidimensional Quality Metrics, MQM; Freitag et al., 2021a), and once with the scorer under evaluation. It then asks how well the scorer’s system ranking agrees with that of the human, as measured by a *meta-metric* such as pairwise accuracy (PA; Kocmi et al., 2021) or soft pairwise accuracy (SPA; Thompson et al., 2024). The WMT metrics shared task (Freitag et al., 2023) performs this meta-evaluation every year, and it sits at the top of the MT stack: it identifies the best scorers, which subsequent research adopts and reports, in turn shaping the next generation of MT systems. Both the scorers and the MT systems therefore silently inherit whatever bias the meta-evaluation imposes.

The MQM ground truth is a multi-component quantity (Freitag et al., 2021a). A translation is penalized both for adequacy errors (failing to convey the meaning of the source) and for fluency errors (being ungrammatical or unnatural). Flamich et al. (2025) show that these two aspects trade off against each other, and Shayegh et al. (2025) show that the tradeoff reaches the meta-evaluation level: the overall human ranking is dominated by whichever aspect varies more across the systems being compared, so the scorers judged best are those aligned with that aspect. In other words, if the systems differ mostly in adequacy, human rankings will be

driven primarily by adequacy as well; consequently, metrics sensitive to adequacy will perform better in the meta-evaluation. Crucially, this lean is heavily driven by system selection (Shayegh et al., 2025).

This is especially concerning because the included systems are a small, filtered subset, thus likely a biased sample of the true, unknown distribution. Figure 1 shows that the lean differs substantially across WMT datasets. Consequently, meta-evaluation produces uncontrolled scorer rankings, risking the endorsement of undesired scorers that misdirect future model development. This motivates the need for an explicit control mechanism.

Shayegh et al. (2025) attempt to address this problem through the lens of “debiasing.” They frame system-selection imbalance as undesirable extrinsic bias and introduce a system-synthesis method to reduce it. However, not all system-selection imbalance is undesirable; only its deviation from the true system distribution is. As the true distribution remains unknown, this attribute is unmeasurable, leading us to abandon debiasing in favor of control. Furthermore, we show that the system synthesis produces unrepresentative systems, yielding a distorted meta-evaluation.

In this work, we expose the previously uncontrolled adequacy–fluency balance as a deliberate, tunable choice. Specifically, we parameterize the evaluation pool using the ratio of weighted marginal variances of adequacy and fluency, a parameter we name β . We treat the systems in the MQM data as a biased sample and reweight them to achieve a target β while minimizing distortion from uniform weighting. Because this reweighting is applied at the system level, the evaluated systems remain real and representative, preserving the meta-evaluation’s original characteristics. This approach is an instance of importance sampling (Tokdar and Kass, 2010) with minimal weight distortion (Deville and Särndal, 1992).

The reweighting is a nontrivial optimization problem, for which we design an exact algorithm, pruning mechanisms that improve efficiency, and certificates that verify optimality. We prove the soundness and completeness of the algorithm, pruning mechanisms, and certificates, and provide an empirical computational-cost analysis.

Validating meta-evaluation is challenging without a ground truth. To address this, we introduce a scorer-augmentation framework that injects controlled signals into real base scorers, preserving the inherent statistical signatures of real scorers while

providing known *relative identity*. Using these augmented scorers, we assess a meta-evaluation’s (1) adequacy-over-fluency preference, (2) MQM adherence, and (3) explainability by MQM. The latter measures noise avoidance while retaining the natural scorer–MQM relationship. Our reweighting method effectively controls the first measure while preserving the other two, outperforming Shayegh et al. (2025)’s synthesis method.

Finally, we report and analyze the meta-evaluation scores of popular scorers, including different variants of xCOMET (Guerreiro et al., 2024) and MetricX (Juraska et al., 2024), across a sweep of the exposed balance parameter β .

In summary, we: (1) formulate the adequacy–fluency balance of MT meta-evaluation as a tunable parameter; (2) provide an exact algorithm that controls this parameter by importance-weighting the systems in the meta-evaluation, minimizing the distortion from uniform weighting so as to maximally preserve the meta-evaluation’s other characteristics; and (3) propose a controlled scorer-augmentation method to assess a meta-evaluation’s adequacy-over-fluency preference, MQM adherence, and explainability by MQM.

2 Background and Related Work

2.1 Meta-Evaluation

A translation scorer is evaluated through meta-evaluation, which measures its agreement with human annotations across a fixed pool of translation systems (Callison-Burch et al., 2007). Expert error annotation under Multidimensional Quality Metrics (MQM; Mariana et al., 2015) has been established as the standard ground truth for the WMT metrics shared task (Freitag et al., 2021a, 2023; Kocmi and Federmann, 2023; Kocmi et al., 2025). Our work focuses on MQM-based meta-evaluation following recent standards.

Meta-metrics. While scorer–human agreement was traditionally measured via Pearson (Macháček and Bojar, 2013, 2014), Mathur et al. (2020) demonstrate that such a correlation is highly sensitive to outlier systems in the evaluation pool. To address this, Kocmi et al. (2021) introduce pairwise accuracy (PA), which measures the fraction of system pairs whose human ordering is correctly reproduced. To overcome PA’s coarse resolution and tie-handling limitations at the segment level (Deutsch et al., 2023), Thompson et al. (2024)

develop soft pairwise accuracy (SPA), which replaces PA’s hard binary comparisons with a continuous measure based on statistical significance. SPA serves as the official system-level meta-metric of WMT24 (Freitag et al., 2024). We follow the line of research and focus on SPA as our main metric. We additionally report results for Pearson correlation in Appendix F.

Meta-evaluation analysis. Recent work critically examines meta-evaluation protocols, studying annotation stability (Knowles, 2021), budget allocation (Riley et al., 2024), test-set item selection (Zouhar et al., 2025), and baseline sensitivity via degenerate sentinel metrics (Perrella et al., 2024). These studies treat pool-induced variation as noise or bias to be detected. Our work treats one special axis of that variation, the adequacy–fluency balance, as a tunable quantity that a practitioner should be able to explicitly configure.

2.2 Adequacy and Fluency

The distinction between translation adequacy and fluency originates in early evaluation campaigns (White et al., 1994) before being largely collapsed into error-weighted penalties (Lommel et al., 2014). Information-theoretic analyses prove that adequacy and fluency genuinely trade off, making a single scalar summary insufficient (Flamich et al., 2025). Moreover, fluency impacts downstream user trust independently of adequacy (Martindale and Carpuat, 2018). This trade-off extends to the meta-evaluation level: meta-evaluation inherits the pool’s lean, favoring adequacy-oriented scorers in recent WMT MQM datasets due to skewed system selection (Shayegh et al., 2025).

We propose a principled parameterization for the adequacy–fluency balance in meta-evaluation, accompanied by a reweighting mechanism to achieve any target balance precisely.

3 Approach

We control the adequacy–fluency balance by reweighting systems while explicitly minimizing the collision probability of the assigned weights. Practically, this keeps the meta-evaluation close to the natural dataset distribution and prevents over-indexing a handful of systems.

To control the adequacy–fluency balance, we need a reference for the two terms. MQM annotations provide fine-grained error categories. We follow Flamich et al. (2025) and group these into

adequacy and fluency subsets, and compute Adequacy MQM and Fluency MQM at the system level accordingly. Their sum, Total MQM, is the meta-evaluation ground truth.³

3.1 The Balance Parameter

A meta-evaluation dataset fixes K translation systems. Write $a=(a_1, \dots, a_K)$ and $f=(f_1, \dots, f_K)$ for their system-level Adequacy MQM and Fluency MQM, so that system k is the point (a_k, f_k) and Total MQM is $t=a+f$. We assign each system k a weight w_k , with $w=(w_1, \dots, w_K)$ normalized so that $\sum_k w_k = \mathbf{1}^\top w = 1$ and $w_k \geq 0$. Reweighting acts on these K weights, so every system in the meta-evaluation remains real; the distribution of systems, which is likely biased, changes.

Let $\Sigma(w) = \text{diag}(w) - ww^\top$. We have weighted mean and variance of scores a as

$$\mu(a; w) = w^\top a \quad (1)$$

$$\begin{aligned} \sigma(a; w)^2 &= a^\top \Sigma(w) a \\ &= \sum_k w_k a_k^2 - \mu(a; w)^2 \end{aligned} \quad (2)$$

respectively, and likewise for f . We also have a – f weighted covariance:

$$\Sigma(a, f; w) = a^\top \Sigma(w) f \quad (3)$$

Setting w to the uniform weighting $\frac{1}{K} \mathbf{1}$ recovers the unweighted sample mean and variance:

$$\mu(a) = \mu(a; \tfrac{1}{K} \mathbf{1}); \quad \sigma(a)^2 = \sigma(a; \tfrac{1}{K} \mathbf{1})^2 \quad (4)$$

We assume throughout that no two systems share the same (a_k, f_k) ; such pairs can be seen as one system in the weighting pool, with the assigned weight uniformly distributed among them.

We define the balance parameter⁴

$$\beta(w) = \frac{\sigma(a; w)^2}{\sigma(a; w)^2 + \sigma(f; w)^2} \in [0, 1] \quad (*)$$

We specifically name the natural balance of the dataset $\beta_0 = \beta(\frac{1}{K} \mathbf{1})$.

Typically, a and f are partially correlated, thus Total MQM’s variance decomposes as $\sigma(t; w)^2 = \sigma(a; w)^2 + \sigma(f; w)^2 + 2 \Sigma(a, f; w)$. Our β uses only the two marginal terms and leaves the cross term $\Sigma(a, f; w)$ out: β answers how much adequacy’s spread contributes relative to fluency’s. In the next part, we show how meta-evaluation responds to this marginal quantity.

³Following Shayegh et al. (2025), we exclude rare errors that fall outside the two categories. The effect is negligible.

⁴An equally valid choice for the balance parameter is $\beta_{\text{std}} = \sigma(a; w) / (\sigma(a; w) + \sigma(f; w))$, which is a monotonic reparametrization $\beta_{\text{std}} = (1 + \sqrt{\beta^{-1} - 1})^{-1}$.

3.2 Meta-Evaluation's Response to β

A meta-metric typically judges scorers by aggregating agreement with the human ranking across system pairs. Therefore, meta-evaluation's lean towards adequacy or fluency is mainly driven by how much each component's gaps vary across the system pairs. We show next that this pairwise-gap variance is equal to the row-level variance, licensing its control through β .

Proposition 1. *For any vector $x=(x_1, \dots, x_K)$ and weights $w=(w_1, \dots, w_K)$ with $\mathbf{1}^\top w = 1$*

$$\sum_{k < k'} w_k w_{k'} (x_k - x_{k'})^2 = \sigma(x; w)^2 \quad (5)$$

Proof.

$$\begin{aligned} & \sum_{k, k'} w_k w_{k'} (x_k - x_{k'})^2 \\ &= \sum_{k, k'} w_k w_{k'} (x_k^2 - 2x_k x_{k'} + x_{k'}^2) \\ &= 2 \sum_k w_k x_k^2 - 2 \left(\sum_k w_k x_k \right)^2 \\ &= 2 \sigma(x; w)^2 \end{aligned} \quad (6)$$

The summand $w_k w_{k'} (x_k - x_{k'})^2$ is symmetric under $k \leftrightarrow k'$ and vanishes when $k=k'$, so splitting the sum by $k < k'$ gives $2 \sum_{k < k'} w_k w_{k'} (x_k - x_{k'})^2$. Dividing (6) by 2 gives the claim. $\square$

Proposition 1 lets us read the row-level variance $\sigma(x; w)^2$ as a pairwise quantity. Now let q be the mean-squared pairwise gap

$$q(x; w) = \frac{\sum_{k < k'} w_k w_{k'} (x_k - x_{k'})^2}{\sum_{k < k'} w_k w_{k'}} \quad (7)$$

By Proposition 1, the numerator in (7) is $\sigma(x; w)^2$ and the denominator is x -independent. Therefore

$$\frac{q(a; w)}{q(a; w) + q(f; w)} = \frac{\sigma(a; w)^2}{\sigma(a; w)^2 + \sigma(f; w)^2} = \beta(w) \quad (8)$$

In other words, β is the balance of mean-squared pairwise gaps, and controlling β effectively drives the meta-evaluation's lean.

3.3 Weighted Soft Pairwise Accuracy

Because SPA is our primary meta-metric, we explicitly define its weighted form. For every system pair (k, k') , SPA compares the confidence that k outranks k' under human judgments against the

confidence under a given scorer. This confidence is estimated via paired-permutation p -values, $p_{kk'}^H$ and $p_{kk'}^m$, the latter for scorer m ; neither depends on w . Classical SPA averages the absolute gap between these p -values uniformly over all $\binom{K}{2}$ pairs. Weighting systems by w assigns a weight of $w_k w_{k'}$ to each pair. Normalizing these weights yields

$$\pi_{kk'} = \frac{w_k w_{k'}}{\sum_{l < l'} w_l w_{l'}} \quad ; \quad \sum_{k < k'} \pi_{kk'} = 1 \quad (9)$$

which defines the weighted meta-metric

$$\text{SPA}(m; w) = 1 - \sum_{k < k'} \pi_{kk'} |p_{kk'}^H - p_{kk'}^m| \quad (10)$$

Uniform weighting yields $\pi_{kk'} = \binom{K}{2}^{-1}$, recovering classical SPA.

3.4 Reweighting as an Optimization Problem

Among all weight vectors w that can realize a given target $\hat{\beta}$, we find the one closest to uniform: it disturbs the natural view least beyond what $\hat{\beta}$ requires, preserving most about the meta-evaluation.

We measure distance from uniform by collision probability $P_{\text{collis}}(w) = \sum_k w_k^2 = w^\top w$. Minimizing this distance directly maximizes effective sample size $\text{ESS}(w) = P_{\text{collis}}(w)^{-1}$, a standard diagnostic for weight concentration. In our work, we report $\text{ESS}(w)/K$ to quantify how broadly the weights are distributed, with K being the size of the original system set.

Writing the balance target as a constraint

$$b(w, \hat{\beta}) = (1 - \hat{\beta})\sigma(a; w)^2 - \hat{\beta}\sigma(f; w)^2 \quad (11)$$

so that $b(w, \hat{\beta}) = 0 \iff \beta(w) = \hat{\beta}$, we formalize the reweighting as an optimization problem:

$$\begin{aligned} w^* = \underset{w}{\operatorname{argmin}} \quad & \frac{1}{2} P_{\text{collis}}(w) \\ \text{s.t.} \quad & \mathbf{1}^\top w = 1, \quad b(w, \hat{\beta}) = 0, \quad w_k \geq 0 \end{aligned} \quad (\mathcal{P})$$

Reachable range. For two systems k, k' , write

$$\beta_{kk'} = \frac{(a_k - a_{k'})^2}{(a_k - a_{k'})^2 + (f_k - f_{k'})^2} \quad (12)$$

and let $\beta_{\min}$ and $\beta_{\max}$ be the smallest and largest such pairwise balance over all pairs of systems in the dataset.

Theorem 1 (Reachable range). *Problem $(\mathcal{P})$ is feasible for target $\hat{\beta}$ iff $\hat{\beta} \in [\beta_{\min}, \beta_{\max}]$. (Proof in Appendix A.1.)*

3.5 Our Algorithm

In this section, we develop and theoretically justify our solver algorithm for Problem (P). We provide a high-level narrative for the algorithm, followed by auxiliary theorems to improve its efficiency.

Define $S(w) = \{k : w_k > 0\}$ as the support system set of w . Call a support S *admissible* for target $\hat{\beta}$ if there exists a feasible weighting w_S with support S for target $\hat{\beta}$, i.e., it is non-negative and $b(w_S, \hat{\beta}) = 0$.

Theorem 2 (Bounded candidate count). *For an admissible support S , let*

$$\begin{aligned} \min_w \frac{1}{2} P_{\text{collis}}(w) \quad (\mathcal{P}_S) \\ \text{s.t. } \mathbf{1}^\top w = 1, \quad b(w, \hat{\beta}) = 0 \quad (\text{no non-negativity}) \end{aligned}$$

be solved on S 's coordinates, with $w_k=0$ for $k \notin S$. Then (a) Problem $(\mathcal{P}_S)$ has at most five stationary points w_S , each obtainable in closed form by solving a polynomial equation of degree ≤ 5 ; and (b) if w^ solves (P) and $S = S(w^*)$, then w^* is one of these w_S . (Proof in Appendix A.2.)*

This gives the abstract shape of our algorithm: for every support S , compute its (at most five) stationary points w_S , discard any with a negative entry, and among everything that survives across every support, return the one of smallest $P_{\text{collis}}(w_S)$. Part (b) guarantees w^* is never missed by this process, thus our algorithm is complete.

Next, we show the enumeration can be largely pruned without losing this guarantee.

For a support S with $|S| \geq 2$, let $\beta_{\min}(S)$ and $\beta_{\max}(S)$ be the smallest and largest pairwise balance $\beta_{kk'}$ among pairs $k, k' \in S$.

Theorem 3 (Admissible supports). *If w_S with support S is a feasible weighting for target $\hat{\beta}$, then $\beta_{\min}(S) \leq \hat{\beta} \leq \beta_{\max}(S)$. (Proof in Appendix A.3.)*

Equivalently, a support whose pairwise balances do not bracket $\hat{\beta}$ can be safely skipped during the search.

Theorem 4 (Support size bound). *For any weighting w supported on S , $P_{\text{collis}}(w) \geq 1/|S|$, with equality iff w is uniform on S . (Proof in Appendix A.4.)*

Theorem 4 suggests that once a feasible weighting with collision probability p has been found, no support of size at most $1/p$ can carry a feasible weighting of strictly smaller collision probability, and such supports may be skipped by the search.

Particularly, our search process visits supports from largest to smallest, and the search stops the first time Theorem 4's pruning applies.

Theorems 3 and 4 offer complementary pruning mechanisms: each is strongest where the other is weakest. When $\hat{\beta}$ is close to the dataset's natural balance β_0 , the optimum stays close to uniform and Theorem 4 prunes heavily. When $\hat{\beta}$ is close to $\beta_{\min}$ or $\beta_{\max}$, it falls out from many supports' admissible range, thus Theorem 3 prunes heavily.

Theorem 5 (Exit certificate). *Let w be feasible for Problem (P). Let $\tilde{a} = a - \mu(a)\mathbf{1}$, $\tilde{f} = f - \mu(f)\mathbf{1}$ be both nonzero, and $\rho = \frac{\tilde{a}^\top \tilde{f}}{\|\tilde{a}\| \|\tilde{f}\|}$ the adequacy-fluency Pearson correlation across systems. If there exist $\eta, \lambda \in \mathbb{R}$ such that*

$$k \in S(w) : \quad w_k + \eta \frac{\partial b(w, \hat{\beta})}{\partial w_k} + \lambda = 0 \quad (\text{i})$$

$$k \notin S(w) : \quad \lambda + \eta \frac{\partial b(w, \hat{\beta})}{\partial w_k} \geq 0 \quad (\text{ii})$$

$$I_2 + 2\eta \begin{pmatrix} \hat{\beta} \|\tilde{f}\|^2 - (1-\hat{\beta})\rho^2 \|\tilde{a}\|^2, & -(1-\hat{\beta})\rho \sqrt{1-\rho^2} \|\tilde{a}\|^2 \\ -(1-\hat{\beta})\rho \sqrt{1-\rho^2} \|\tilde{a}\|^2, & -(1-\hat{\beta})(1-\rho^2) \|\tilde{a}\|^2 \end{pmatrix} \succeq 0 \quad (\text{iii})$$

where $\succeq 0$ means positive semidefinite, then w is a global optimum of Problem (P). (Proof in Appendix A.5.)

Theorem 5 grants our algorithm a short exit: the moment one stationary point on a visited support passes the three conditions, the search stops.

Algorithm 1 concludes our exact search algorithm. Theorems 1–5 guarantee the soundness and completeness of our algorithm unconditionally: pruned branches are guaranteed to not include better solutions, short exit guarantees returning the global optimum, and the algorithm guarantees to search all branches (finite) carrying a possible solution. We also cross-validated its output against brute-force exhaustive search over all supports on small system sets, where full enumeration remains tractable, and found exact agreement throughout.⁵ In Appendix B, we report how much computation Theorems 3–5 save us in practice.

4 Experiments

In this section, we analyze the response of meta-evaluation to different leanings that scor-

⁵Theorem 2 bounds the stationary points by the roots of a degree-five polynomial, but our implementation solves the equivalent rational secular equation directly rather than rooting that polynomial, whose coefficients become badly scaled as $\theta_2 \rightarrow 0$. This and all numerical thresholds are fixed in the released code.

Algorithm 1 Exact reweighting for target $\hat{\beta}$

Require: $\hat{\beta} \in [\beta_{\min}, \beta_{\max}]$ {Theorem 1}

```
1: register  $w^*$ 
2: for  $S \subseteq \{1, \dots, K\}$ ,  $|S| \geq 2$ , in decreasing
   order of  $|S|$  do
3:   if  $|S| \leq 1/P_{\text{collis}}(w^*)$  then
4:     break {Theorem 4}
5:   end if
6:   if  $S$  not admissible at  $\hat{\beta}$  then
7:     continue {Theorem 3}
8:   end if
9:   for each stationary candidate  $w$  on  $S$  {The-
      orem 2a} do
10:    if  $w$  passes the exit certificate then
11:      return  $w$  {Theorem 5}
12:    end if
13:    if  $w$  feasible and  $P_{\text{collis}}(w) < P_{\text{collis}}(w^*)$ 
       then
14:       $w^* \leftarrow w$  {Theorem 2b}
15:    end if
16:  end for
17: end for
18: return  $w^*$ 
```

ers may have. Specifically, we measure each meta-evaluation’s adequacy-over-fluency preference, MQM adherence, and explainability by MQM. To this end, we develop a controlled testbed that reveals signals for these aspects. Then, we benchmark our reweighting mechanism on this testbed, comparing it against the unweighted meta-evaluation and the system-synthesis-based meta-evaluation of [Shayegh et al. \(2025\)](#).

4.1 Scorer Augmentation

Evaluation and analysis of scorers needs a ground truth for their true underlying character (e.g., how much of a scorer’s assessment is driven by the fluency of the translation), which is not available. To obtain a replacement, we propose to augment scorers in a controlled way, so that the augmented scorers’ relative identities are known by construction, even though no scorer’s absolute identity is. We use this relative identity as pseudo ground truth for evaluation and analysis. Importantly, we augment real scorers, staying close to their natural biases and characteristics, except for those we control in particular.

We augment a base scorer s_0 in three directions, giving three families of augmented scorers, each to

test a different way a meta-evaluation’s preference among scorers could change. Given a and f being gold Adequacy MQM and Fluency MQM, the three directions are:

- MQM-adherence family: $s_0 \leftrightarrow a + f$
- Adequacy–fluency family: $s_0 \leftrightarrow a - f$
- Explainability-by-MQM family: $s_0 \leftrightarrow$ part of s_0 only explained by (a, f)

An ideal meta-evaluation should typically prefer an augmented scorer that better adheres to MQM or prefers explainability by MQM.

We realize each family with a dial d : a base scorer always has $d = 0$, and as d increases, the augmented scorer approaches its family’s target. Each value of d yields a new augmented scorer whose relative identity to other members of its family is known through their relative d values. Note that we allow negative d , i.e., going backwards in each direction.

MQM-adherence family. This is constructed by moving linearly towards $t = a + f$ at the system level. For segment i of system k , and base scorer s_0 , the augmented scorer is

$$s_d(k, i) = s_0(k, i) + \frac{d \cdot \sigma(s_0)}{\sigma(z)} z_k \quad (13)$$

where z is the system-level z -score of t , and σ is standard deviation across systems. In simple words, we move towards z in steps of $d \times s_0$ ’s scale.⁶

The above construction conserves characteristics of real scorers: the two ends of the direction ($d = 0$ and $d \rightarrow \infty$) are the base scorer itself and the MQM ground truth, both considered valid scorers. Moreover, for every finite d , the augmented scorer preserves the base scorer’s own system- and segment-specific (k, i -related) biases and patterns. Note that the above construction does keep the within-system variance $\sigma_i^2(s_d(k, \cdot))$ of each system k intact. This is the signal SPA reads as a scorer’s confidence. Keeping it unchanged avoids confounding the MQM-adherence signal with a change in SPA’s confidence signal.

⁶The same approach can produce Adequacy MQM-adherence or Fluency MQM-adherence families of synthetic scorers by considering z as the system-level z -score of a or f , respectively. We exclude these families from our analysis as the partial correlation between the two aspects ([Shayegh et al., 2025](#)) confounds their interpretation.

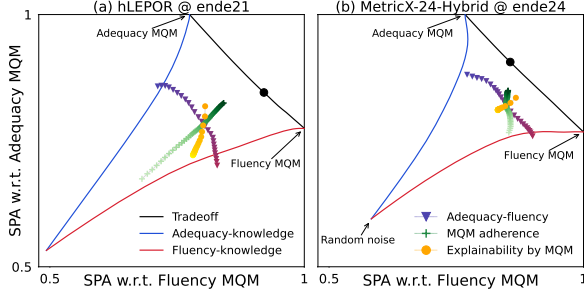


Figure 2: The SPA plane of Shayegh et al. (2025), with the augmented scorers of all three families plotted as points, for (a) hLEPOR (Han et al., 2013) in ende21 dataset and (b) MetricX-24-Hybrid (Juraska et al., 2024) in ende24 dataset as the base scorer.

Adequacy–fluency family. Similar to MQM-adherence family, we write

$$s_d(k, i) = s_0(k, i) + \frac{d \cdot \sigma(s_0)}{\sigma(z^{(a)})} z_k^{(a)} - \frac{d \cdot \sigma(s_0)}{\sigma(z^{(f)})} z_k^{(f)} \quad (14)$$

where $z^{(a)}$ and $z^{(f)}$ are the system-level z -scores of a and f , respectively, and σ is the standard deviation across systems. This construction is particularly clean: with correlation $\rho(a, f) < 1$, it establishes a known adequacy-over-fluency preference along the increasing dial d . If $\rho = 1$, there is no trade-off, and the construction has no effect.

Explainability-by-MQM family. This family’s direction is given by the part of the base scorer not explained by the joint (a, f) pair: moving one way along it reduces this unexplained part, approaching what (a, f) alone explains, while moving the other way amplifies it. It is different from the other two as instead of injecting/removing specific signals, it magnifies/reduces any (a, f) -unrelated signal the base scorer may have. We find the direction as follows. Write a_{ki} and f_{ki} for segment i of system k ’s gold Adequacy MQM and Fluency MQM scores, and let $\bar{s}_0(a, f)$ be the s_0 ’s mean score among segments sharing the exact (a, f) :

$$\bar{s}_0(a, f) = \mathbb{E}_{i, k: a_{ki}=a, f_{ki}=f} [s_0(k, i)] \quad (15)$$

Let

$$\bar{e}_k = \mathbb{E}_i [s_0(k, i) - \bar{s}_0(a_{ki}, f_{ki})] \quad (16)$$

be system k ’s *offset*: the part of its score that the joint pair does not explain. We construct the augmented scorer

$$s_d(k, i) = s_0(k, i) - d \cdot \bar{e}_k. \quad (17)$$

In simple words, we partially remove each system’s offset beyond what its (a, f) explains. For negative d , we add more of it.

To visually illustrate the three families, Figure 2 plots them, applied to two different base scorers, on the SPA plane of Shayegh et al. (2025).

4.2 Evaluation Setup

We analyze the performance of a meta-evaluation separately in each of three directions. Specifically, we compare the meta-evaluation’s assessment of different augmented scorers within each family, and count for how many adjacent-in- d scorer pairs the meta-evaluation correctly identifies the dial direction. We report this success rate separately for each family as three measures.

Datasets. We use MQM data published by Freitag et al. (2021a).⁷ We restrict our analysis to datasets from 2021 onward (Freitag et al., 2021b, 2022, 2023, 2024) due to limited standardization and support for older datasets. We specifically exclude ende23 because many of its MQM annotations are categorized as “other” rather than being properly assigned to adequacy or fluency categories. This leaves 11 datasets: ende24, enes24, jazh24, heen23, zhen23, ende22, zhen22, ende21, zhen21, ted_ende, and ted_zhen. Across all datasets, we evaluate all available systems except for M2M100_1.2B-B4 and MSLC, which are officially flagged for exclusion. Detailed statistics are in Appendix D.

Base scorers. We consider automatic scorers (omitting human references) in the MQM datasets as the base scorers. This includes the competitive scorers as well as the diagnostic ones included by WMT. We exclude those without segment-level scores (e.g., corpus-level chrF). We also deduplicate perfectly duplicated scorer pairs. We provide the complete list of the base scorers in Appendix D. For each meta-evaluation and each dataset, we report the average scores across all base scorers.

Dial grid. The closer two d values are, the harder it is for a meta-evaluation to recognize their relative identity. We empirically choose the d grid for the best readability of results. For MQM adherence, d takes $-1.5 + 3^{n/200}$, and for explainability by MQM, $-3.5 + 4^{n/200}$, where $n \in \mathbb{N}$:

⁷github.com/google/wmt-mqm-human-evaluation; MQM data for 2025 (Lavie et al., 2025) has not been published by the time of this study.

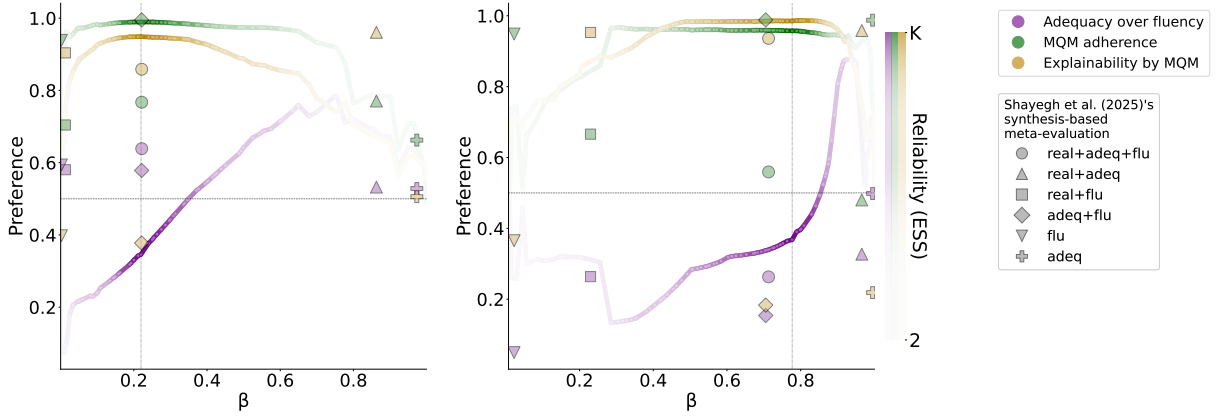


Figure 3: Main results on ende21 (left) and ende24 (right). Curve color intensity represents the effective sample size (ESS); a lower ESS indicates fewer actively involved systems, reducing representativeness. The strictly increasing adequacy-over-fluency curve demonstrates that β is an effective driver of this balance. The natural balance β_0 for each dataset is marked by a vertical dotted line, with its intersections indicating the original meta-evaluation’s performance. Higher MQM adherence and explainability by MQM indicate better meta-evaluation internal consistency. Markers show the performance of the six synthesis-based meta-evaluation setups from [Shayegh et al. \(2025\)](#), placed at their synthesized datasets’ respective β_0 .

$0 \leq n \leq 200$, resulting in 200 adjacent pairs per base scorer. For adequacy-over-fluency preference, d ranges from -0.5 to 0.5 with a step size of 0.02 , resulting in 50 adjacent pairs per base scorer.

4.3 Main Results

Figures⁸ 3 and 7 show our method’s performance across datasets over a β sweep spanning $[\beta_{\min}, \beta_{\max}]$, plus explicitly every $\beta_{kk'}$ in the dataset. The monotonically increasing adequacy-over-fluency curves confirm that our reweighting mechanism effectively controls this balance. Furthermore, within a range around β_0 where ESS remains large, our method successfully preserves the original meta-evaluation’s MQM adherence and explainability by MQM.

The strong correlation between ESS and both MQM adherence and explainability by MQM confirms ESS as a useful proxy for reliability, preventing severe evaluation distortion. Practically, ESS identifies the reachable range for controlling the adequacy–fluency balance without degrading meta-evaluation quality.

The permissible deviation from β_0 is fundamentally constrained by the dataset itself. In datasets like jazh24, most $\beta_{kk'}$ values concentrate within a narrow band, limiting reweighting capacity. Consequently, if a dataset lacks systems that intrinsically support a fluency-heavy evaluation, this imposes a strict limit on the achievable balance.

⁸Results for some datasets are deferred to the Appendix E due to space constraints. All the claims are written with respect to all datasets, including those deferred to the appendix.

Compared to [Shayegh et al. \(2025\)](#), our method consistently yields superior MQM adherence and/or explainability by MQM across all datasets, whether controlled by β or by adequacy-over-fluency preference. The only exceptions occur at extreme adequacy-over-fluency values (e.g., in zhen23, ende22, and ende24) which their synthetic method reaches but our reweighting cannot. However, explainability by MQM heavily drops for both methods in these regimes, rendering those extremes highly distorted and practically infeasible anyway. Overall, our method strictly improves upon [Shayegh et al. \(2025\)](#) in preserving the internal consistency of the meta-evaluation.

Analyzing the markers in Figure 3 reveals a critical limitation of synthetic systems. In [Shayegh et al. \(2025\)](#)’s setups that retain real systems, explainability by MQM is preserved but MQM adherence degrades. Conversely, in their fully synthesized pools, MQM adherence is preserved while explainability by MQM collapses. Explainability by MQM measures the meta-evaluation’s response to scorer noise while preserving natural biases; it is informed, but not strictly dictated, by MQM. This failure point of fully synthetic pools confirms our hypothesis that synthetic systems are unrepresentative. This stems from their construction: segments are regrouped to maximize between-system variance. By the law of total variance ([Soch et al., 2025](#)), since total pool variance is fixed, altering the between-system component inevitably alters the within-system component. This yields systems

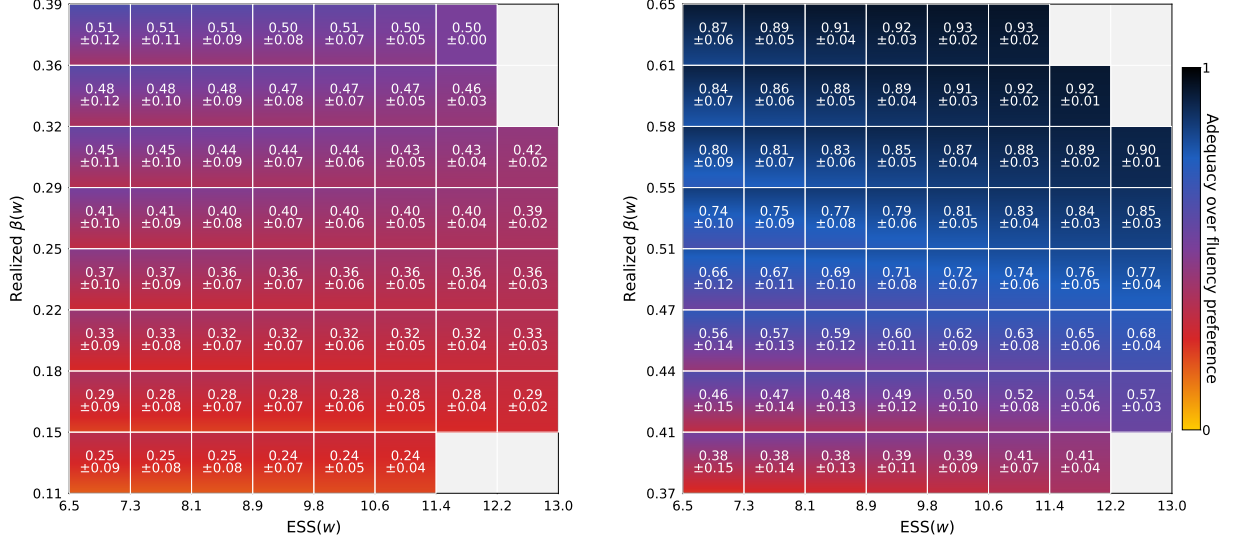


Figure 4: Adequacy-over-fluency preference across the (ESS, β) plane for ende21 (left) and ted_zhen (right). Each cell gives the mean and standard deviation over the weightings realizing that (ESS, β), color-ramped from mean-std at its bottom edge to mean+std at its top. Grey cells are unreachable.

with unrealistically low score variances, being unrepresentative and also distorting SPA’s confidence estimations.

In Appendix F, we report similar experiments using Pearson correlation as the meta-metric. The observed trends remain consistent, further supporting the effectiveness of our reweighting method.

4.4 Additional Analysis

Variance Analysis. We investigate whether weight vectors with similar β values can yield substantially different adequacy-fluency balances due to uncontrolled factors, which would invalidate β as a reliable control parameter. The results in §4.3 are limited to P_{collis} -minimized weight vectors for each β , hiding any same- β variance. To address this, we sample numerous weight vectors on the simplex $w \geq 0 : \mathbf{1}^\top w = 1$, each realizing some $\beta(w)$, ESS(w), and adequacy-over-fluency preference. By grouping these into buckets of (ESS, β) pairs, we compare within-bucket preference variances (indicating β -independent shifts) against across-bucket average trends (indicating the β -effect), demonstrating the reliability of β as a control parameter.

We restrict the experiment to $w : \text{ESS}(w) \in [\frac{K}{2}, K]$, given the unreliability of weighting with lower ESS. We draw around 3.5M weight vectors per dataset. These sets provide varying β coverage across datasets. We partition β and ESS ranges into 8 equal-sized buckets each. We compute the adequacy-over-fluency preference for each weighting, presenting the results as a heatmap over this

8×8 grid. Because these weightings are unoptimized, they remain unconstrained beyond the β and ESS. Consequently, a cell’s standard deviation strictly isolates the variation in SPA-based adequacy-over-fluency preference attributable to everything about w except β .

Figures⁹ 4 and 9 report the results. Each cell represents a pair of β and ESS buckets. Reading down any column, preference increases with β at every ESS. This trend holds for 99.1% of vertically-adjacent cell pairs pooled across all datasets. Reading across rows, higher ESS sharply reduces variance: averaged over all 554 horizontally-adjacent cell pairs, higher ESS exhibits 18.3% lower same- β standard deviation, and the standard deviation decreases in 99.8% of those pairs. Lastly, comparing means to standard deviations confirms that same- β variance does not obscure the β signal. Despite partial overlap between cells, the expected differences remain substantial. A one-sided Welch’s t -test (Welch, 1947) on every vertically-adjacent cell pair confirms this: 94.8% of pairs show that the higher- β cell has a significantly higher preference ($p < 0.01$), and no pair is significant in the opposite direction.

The mapping from β to the adequacy-over-fluency preference is dataset-dependent: not only does the preference measure preclude cross-dataset comparison by construction, identical β values also yield empirically different preference scores across

⁹Results for some datasets are deferred to Appendix G due to space limits. Claims are made with respect to all datasets.

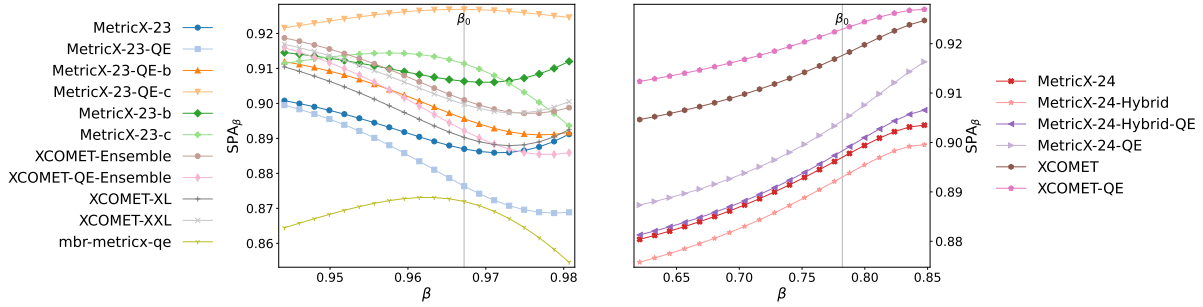


Figure 5: Weighted meta-evaluation performance (SPA_β) of MetricX and xCOMET variants as a function of the adequacy–fluency balance parameter β . In zhen23 (left), scorer rankings change depending on the target balance. In ende24 (right), relative performance remains stable across the reachable range. β interval is bounded by $ESS \geq K - 1$, where K is the total number of systems (15 for zhen23, 16 for ende24).

datasets. Despite this limitation, the underlying relationship between β and the adequacy–fluency balance remains universal. Its direction, monotonicity, and robustness to same- β variation hold across every dataset tested, licensing β as a reliable control parameter.

Weighted Meta-Evaluation. Having validated our reweighting mechanism, we now apply it to compare variants of the popular scorers MetricX (Juraska et al., 2024) and xCOMET (Guerreiro et al., 2024) across a sweep of β . Figure 5 reports the weighted soft pairwise accuracy (SPA) for the zhen23 and ende24 datasets.¹⁰

The rankings in ende24 remain largely stable across the reachable β range. In contrast, the performance trajectories in zhen23 intersect a lot, demonstrating that the relative ranking of scorers **can** be highly sensitive to the adequacy–fluency balance. Our proposed parameter and reweighting tool explicitly describe and control this behavior.

More results in the appendix. In Appendix B, we empirically evaluate the computational efficiency of our search algorithm, ablate its pruning mechanisms, and compare it against a numerical optimization baseline. In Appendix C, we show that meta-evaluation is sensitive to random system removal, and that adequacy–fluency imbalance is not the main cause of this instability.

5 Conclusion

Machine translation meta-evaluation inherits the adequacy–fluency tradeoff, with its balance dictated by the specific combination of translation systems in the meta-evaluation data. Because this system

set is a small, unrepresentative sample that fluctuates across datasets, the resulting adequacy–fluency balance is uncontrolled and subject to change.

We exposed this balance as a tunable control parameter, β . To achieve a target β , we reweighted existing systems while minimizing distortion from uniform weighting, ensuring the evaluation pool remains real and representative. We supported this with an exact optimization algorithm, equipped with theoretical guarantees and pruning mechanisms, to efficiently compute these weights.

To provide an internal diagnostic for meta-evaluation behavior, we introduced a scorer-augmentation framework that establishes a known relative identity for near-real scorers. Our experiments demonstrated that our reweighting method successfully controls the adequacy–fluency balance and preserves evaluation integrity, outperforming system-synthesis approaches. Finally, our analysis of popular scorers showed that scorer rankings can be sensitive to β .

Calibration of β in practice. A specific numerical value of β does not yield an identical adequacy-over-fluency preference across different datasets. This stems from inherent variations in language pairs, the correlation between Adequacy MQM and Fluency MQM, base scorer biases, and other unaccounted variables. Thus, β functions as a directional control knob rather than an absolute metric. To determine the target β for a specific real-world scenario (e.g., technical versus creative translation), we recommend practitioners craft a few translation pairs that exhibit a marginal adequacy–fluency tradeoff, where one translation is slightly more adequate and the other slightly more fluent. After establishing which direction their specific use case prefers, they can sweep β until the meta-evaluation prioritizes scorers aligned with that preference.

¹⁰We selected the two datasets empirically to best show two different possible scenarios: when the scorer ranking is or is not influenced by adequacy–fluency balance.

Limitations

Our reweighting mechanism is strictly bounded by the intrinsic diversity of the underlying dataset. If a dataset lacks systems that exhibit isolated adequacy or fluency errors (e.g., jazh24), the reachable range of β is narrow. Forcing β beyond this range requires sparse weights, which reduces the effective sample size (ESS) and compromises the representativeness of the evaluation pool. Further expanding the admissible range for those datasets with limited flexibility remains a direction for future work.

Moreover, our framework depends on the provided MQM annotations and the chosen adequacy–fluency taxonomy, introducing two unaccounted sources of variance. First, WMT human labels contain noise and individual annotator bias. Our method does not model this uncertainty, meaning it propagates directly into system-level scores and β . Second, categorizing errors as adequacy or fluency is subjective. A different taxonomy alters the score distributions, so the same numerical β will behave differently under a different decomposition. Accounting for annotation uncertainty and taxonomy dependence requires further research.

Finally, controlling the adequacy–fluency balance does not resolve the broader instability of meta-evaluation. As demonstrated in Appendix C, scorer rankings remain sensitive to system selection even when β is fixed. Future work must identify the root causes of system-selection instability. We would like to investigate other axes of variation across translation systems that disrupt scorer rankings.

Acknowledgements

We thank Lili Mou for the insightful discussions that provided the initial motivation for this research, and Jan-Thorsten Peter for their valuable feedback on the final manuscript. We also extend our gratitude to the WMT reviewers, both for this manuscript and for our prior work last year, whose constructive feedback significantly shaped the direction of this study. This research was conducted independently and received no external funding.

References

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.

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.

Jean-Claude Deville and Carl-Erik Särndal. 1992. Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418):376–382.

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, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81.

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 (WMT)*, 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.

Ramanathan Gnanadesikan and John R. Kettenring. 1972. Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, 28(1):81–124.

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.

- Aaron Li-Feng Han, Derek F. Wong, Lidia S. Chao, Liangye He, Yi Lu, Junwen Xing, and Xiaodong Zeng. 2013. [Language-independent model for machine translation evaluation with reinforced factors](#). In *Proceedings of Machine Translation Summit XIV: Posters*.
- 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.
- Maurice G Kendall et al. 1938. A new measure of rank correlation. *Biometrika*, 30:81–93.
- Rebecca Knowles. 2021. [On the stability of system rankings at WMT](#). In *Proceedings of the Sixth Conference on Machine Translation*, pages 464–477.
- Tom Kocmi, Sweta Agrawal, Ekaterina Artemova, Eleftherios Avramidis, Eleftheria Briakou, Pinzhen Chen, Marzieh Fadaee, Markus Freitag, Roman Grundkiewicz, Yupeng Hou, Philipp Koehn, Julia Kreutzer, Saab Mansour, Stefano Perrella, Lorenzo Proietti, Parker Riley, Eduardo Sánchez, Patricia Schmidtova, Mariya Shmatova, and Vilém Zouhar. 2025. [Findings of the WMT25 multilingual instruction shared task: Persistent hurdles in reasoning, generation, and evaluation](#). In *Proceedings of the Tenth Conference on Machine Translation*.
- Tom Kocmi and Christian Federmann. 2023. [Large language models are state-of-the-art evaluators of translation quality](#). In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203.
- 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, Greg Hanneman, Sweta Agrawal, Diptesh Kanojia, Chi-Kiu Lo, Vilém Zouhar, Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, Archchana Sindhuja, Jiayi Wang, David Ifeoluwa Adelani, Brian Thompson, Tom Kocmi, Markus Freitag, and Daniel Deutsch. 2025. [Findings of the WMT25 shared task on automated translation evaluation systems: Linguistic diversity is challenging and references still help](#). In *Proceedings of the Tenth Conference on Machine Translation*.
- Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. [Multidimensional quality metrics \(MQM\): A framework for declaring and describing translation quality metrics](#). *Tradumàtica: tecnologies de la traducció*, pages 455–463.
- 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.
- Matouš Macháček and Ondřej Bojar. 2014. [Results of the WMT14 metrics shared task](#). In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 293–301. Association for Computational Linguistics.
- 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, Timothy Baldwin, and Trevor Cohn. 2020. [Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997.
- Jorge J Moré and Danny C Sorensen. 1983. Computing a trust region step. *SIAM Journal on scientific and statistical computing*, 4(3):553–572.
- Jorge Nocedal and Stephen J Wright. 2006. *Numerical optimization*. Springer.
- Jaehyun Park and Stephen Boyd. 2017. General heuristics for nonconvex quadratically constrained quadratic programming. *arXiv preprint arXiv:1703.07870*.
- 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.
- 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.
- Behzad Shayegh, Jan-Thorsten Peter, David Vilar, Tobias Domhan, Juraj Juraska, Markus Freitag, and Lili Mou. 2025. [Feeding two birds or favoring one? adequacy–fluency tradeoffs in evaluation and meta-evaluation of machine translation](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 269–285.
- Joram Soch, Karahan Sarıtaş, Pietro Monticone, Thomas J Faulkenberry, Osvaldo A Martin, Alex Kipnis, Salvador Balkus, Carsten Allefeld, Heiner

Atze, Adam Knapp, et al. 2025. Statproofbook/statproofbook. github.io: Statproofbook 2024. Zenodo.

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.

Surya T Tokdar and Robert E Kass. 2010. Importance sampling: a review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2:54–60.

B. L. Welch. 1947. [The generalization of ‘student’s’ problem when several different population variances are involved](#). *Biometrika*, 34(1/2):28–35.

John S. White, Theresa O’Connell, and Francis O’Mara. 1994. The ARPA MT evaluation methodologies: Evolution, lessons, and future approaches. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, Columbia, Maryland, USA.

Hsien-Chung Wu. 2007. [The karush–kuhn–tucker optimality conditions in an optimization problem with interval-valued objective function](#). *European Journal of Operational Research*, 176:46–59.

Vilém Zouhar, Peng Cui, and Mrinmaya Sachan. 2025. [How to select datapoints for efficient human evaluation of NLG models?](#) *Transactions of the Association for Computational Linguistics*.

A Proofs

A.1 Proof of Theorem 1

Theorem 1 (Reachable range). *Problem $(\mathcal{P})$ is feasible for target $\hat{\beta}$ iff $\hat{\beta} \in [\beta_{\min}, \beta_{\max}]$.*

Proof. The forward direction is given by Theorem 3 applied to the full support. We now prove the reverse direction: Any $\hat{\beta} \in [\beta_{\min}, \beta_{\max}]$ is a feasible target for Problem $\mathcal{P}$.

$\beta(w)$ is continuous on the relative interior of S ’s simplex face and that face is connected. Moreover, for two systems $k_1, k_2 : \beta_{\min}(S) = \beta_{k_1 k_2}$, a weight vector $\hat{w} = \{k \neq k_1, k_2 : w_k = 0\}$ drives $\beta(w) = \beta_{\min}(S)$; symmetrically for $\beta_{\max}(S)$. So the image is a continuous interval with infimum $\beta_{\min}(S)$ and supremum $\beta_{\max}(S)$. $\square$

A.2 Proof of Theorem 2

Theorem 2 (Bounded candidate count). *For an admissible support S , let*

$$\begin{aligned} \min_w \frac{1}{2} P_{\text{collis}}(w) & \quad (\mathcal{P}_S) \\ \text{s.t. } \mathbf{1}^\top w = 1, \quad b(w, \hat{\beta}) = 0 & \quad (\text{no non-negativity}) \end{aligned}$$

be solved on S ’s coordinates, with $w_k = 0$ for $k \notin S$. Then (a) Problem $(\mathcal{P}_S)$ has at most five stationary points w_S , each obtainable in closed form by solving a polynomial equation of degree ≤ 5 ; and (b) if w^ solves $(\mathcal{P})$ and $S = S(w^*)$, then w^* is one of these w_S .*

Proof. [Part (a)] We prove this part by deriving the explicit polynomial to bound the stationary points. By (11) and Proposition 1:

$$\begin{aligned} b(w, \hat{\beta}) &= (1 - \hat{\beta}) \sum_{k < k'} w_k w_{k'} (a_k - a_{k'})^2 \\ &\quad - \hat{\beta} \sum_{k < k'} w_k w_{k'} (f_k - f_{k'})^2 \end{aligned} \quad (18)$$

$$\begin{aligned} &= (1 - \hat{\beta}) \left(\sum_k w_k a_k^2 - w^\top a a^\top w \right) \\ &\quad - \hat{\beta} \left(\sum_k w_k f_k^2 - w^\top f f^\top w \right) \end{aligned} \quad (19)$$

$$\begin{aligned} &= w^\top (\hat{\beta} f f^\top - (1 - \hat{\beta}) a a^\top) w \\ &\quad + \sum_k w_k ((1 - \hat{\beta}) a_k^2 - \hat{\beta} f_k^2) \end{aligned} \quad (20)$$

$$= w^\top H w + w^\top g \quad (21)$$

where $H = \hat{\beta} f f^\top - (1 - \hat{\beta}) a a^\top$ and $g_k = (1 - \hat{\beta}) a_k^2 - \hat{\beta} f_k^2$, yielding a quadratic in w .

Eliminate the budget constraint $\mathbf{1}^\top w = 1$ using the affine transformation

$$w = u_0 + N u \quad (22)$$

where u_0 is the uniform weighting on S and N has orthonormal columns spanning $\{v : \mathbf{1}^\top v = 0\}$. Substituting (22) into (21) yields a single quadratic in the free variable u :

$$b(w, \hat{\beta}) = \tilde{b}(u, \hat{\beta}) = u^\top M u + u^\top r + c = 0 \quad (23)$$

where $M = N^\top H N$, $r = N^\top (2H u_0 + g)$, and $c = b(u_0, \hat{\beta})$. We can restate the problem in the free variable u :

$$\min_u \frac{1}{2} \|u\|^2 + \text{const.} \quad \text{s.t. } \tilde{b}(u, \hat{\beta}) = 0 \quad (\mathcal{P}_S^{(u)})$$

A point u is a *stationary point* of $(\mathcal{P}_S^{(u)})$ if it is feasible and satisfies

$$\nabla_u \left(\frac{1}{2} \|u\|^2 + \eta \cdot \tilde{b}(u, \hat{\beta}) \right) = 0 \quad (24)$$

for a multiplier $\eta \in \mathbb{R}$. Taking the derivative:

$$u + \eta(2M u + r) = 0 \quad (25)$$

$$(I + 2\eta M) u = -\eta r \quad (26)$$

We call the support S *degenerate* if all its pairwise balances $\beta_{kk'}$, $k, k' \in S$, coincide. We consider two scenarios:

[Case 1: S is degenerate] Let $\bar{\beta}$ be the common value of the pairwise balances in S . Because S is admissible, Theorem 3 forces $\hat{\beta} = \beta_{\min}(S) = \beta_{\max}(S) = \bar{\beta}$. For any pair $k, k' \in S$ we have $(a_k - a_{k'})^2 + (f_k - f_{k'})^2 > 0$, as no two systems share the same (a_k, f_k) ; together with $\hat{\beta} = \beta_{kk'}$ and (12) this gives

$$(1 - \hat{\beta})(a_k - a_{k'})^2 - \hat{\beta}(f_k - f_{k'})^2 = 0 \quad (27)$$

Substituting into (18)

$$\begin{aligned} b(w, \hat{\beta}) &= \sum_{k < k'} w_k w_{k'} ((1 - \hat{\beta})(a_k - a_{k'})^2 - \hat{\beta}(f_k - f_{k'})^2) \\ &= \sum_{k < k'} w_k w_{k'} \cdot 0 = 0 \end{aligned} \quad (28)$$

for every w . So the quadratic $\tilde{b}(\cdot, \hat{\beta})$ vanishes identically and $M = 0$, $r = 0$, $c = 0$ in (23). The stationarity condition (26) reads $u = 0$, so S carries exactly one stationary point: the uniform weighting u_0 . It is feasible, minimizes the strictly convex objective $\frac{1}{2}\|u\|^2$, and is the unique global optimum on S .

[Case 2: S is non-degenerate] The pairwise balances in S are not all equal. Let $\tilde{a}$ and $\tilde{f}$ be the score vectors restricted to S and centered by their uniform means over S :

$$\tilde{a}_k = a_k - \mu(a; u_0) \quad (29)$$

$$\tilde{f}_k = f_k - \mu(f; u_0) \quad (30)$$

for $k \in S$. These are linearly independent: were $\tilde{f} = \kappa \tilde{a}$ for some $\kappa \in \mathbb{R}$, every pair would satisfy $(f_k - f_{k'}) = \kappa(a_k - a_{k'})$ and hence $\beta_{kk'} = (1 + \kappa^2)^{-1}$, contradicting non-degeneracy; the same holds with the roles of $\tilde{a}$ and $\tilde{f}$ exchanged.

Within the zero-sum subspace defined by N , M acts as the difference of two independent rank-1 outer products: $\hat{\beta} \tilde{f} \tilde{f}^\top$ and $(1 - \hat{\beta}) \tilde{a} \tilde{a}^\top$. By definition of β , we have $0 < \hat{\beta} < 1$. Therefore, this difference provides one strictly positive and one strictly negative direction. Consequently, M is indefinite with rank two, possessing exactly two nonzero eigenvalues:

$$\theta_1 > 0 > \theta_2 \quad (31)$$

We use this structure to solve for the stationary candidates. Projecting (26) onto the eigen basis $\{v_j\}_{j=1}^{|S|-1}$ of M , and defining the coordinate components $u_j = v_j^\top u$ and $r_j = v_j^\top r$, splits it into independent scalar equations of the form

$$u_j(1 + 2\eta\theta_j) = -\eta r_j \quad (32)$$

The system is singular exactly at the two poles $\lambda_j = -1/(2\theta_j)$ for $j \in \{1, 2\}$. We categorize the candidates by resolving across three subcases for η :

[Case 2.1: $\eta = \lambda_j$ and $r_j \neq 0$] The j -th equation reads $u_j \cdot 0 = -\lambda_j r_j \neq 0$. The system is inconsistent, yielding zero candidates.

[Case 2.2: $\eta = \lambda_j$ and $r_j = 0$] The j -th equation is $0 = 0$, leaving the coordinate u_j free. For every other coordinate $m \neq j$, the coefficient $(1 + 2\lambda_j \theta_m) = 1 - \frac{\theta_m}{\theta_j}$ is strictly non-zero: $\theta_m = 0$ for flat directions $m \geq 3$ and $\frac{\theta_{3-j}}{\theta_j} < 0$ by (31). Therefore, every $u_m : m \neq j$ is uniquely fixed to

$$u_m = \frac{-\lambda_j r_m}{1 + 2\lambda_j \theta_m} \quad (33)$$

Defining u° as this baseline vector with $u_j^\circ = 0$, the full quadratic constraint separates into:

$$\tilde{b}(u, \hat{\beta}) = \theta_j u_j^2 + r_j u_j + \tilde{b}(u^\circ, \hat{\beta}) = 0 \quad (34)$$

Because $r_j = 0$, the linear term vanishes, leaving $\theta_j u_j^2 = -\tilde{b}(u^\circ, \hat{\beta})$. This yields at most two candidates:

$$u_j = \pm \sqrt{-\tilde{b}(u^\circ, \hat{\beta})/\theta_j} \quad (35)$$

[Case 2.3: $\eta \neq \lambda_j$] In this case, $1 + 2\eta\theta_j \neq 0$, which gives

$$u_j(\eta) = \frac{-\eta r_j}{1 + 2\eta\theta_j} \quad (36)$$

which simplifies to $-\eta r_j$ for $j \geq 3$. We substitute these components into (23).

When $j \geq 3$, the quadratic terms vanish by $\theta_j = 0$, and the linear terms sum to:

$$\sum_{j \geq 3} r_j (-\eta r_j) = -\eta \sum_{j \geq 3} r_j^2 \quad (37)$$

For $j \in \{1, 2\}$, substituting $u_j(\eta)$ into its quadratic and linear terms yields:

$$\begin{aligned} &\theta_j u_j(\eta)^2 + r_j u_j(\eta) \\ &= \theta_j \left(\frac{-\eta r_j}{1 + 2\eta\theta_j} \right)^2 + r_j \left(\frac{-\eta r_j}{1 + 2\eta\theta_j} \right) \end{aligned} \quad (38)$$

$$= -\eta r_j^2 \frac{(1 + \eta\theta_j)}{(1 + 2\eta\theta_j)^2} \quad (39)$$

Combining the constant c , (37), and (39) for the two j values, (23) becomes:

$$c - \eta \left[\sum_{j=1,2} \frac{r_j^2(1 + \theta_j \eta)}{(1 + 2\eta\theta_j)^2} + \sum_{j \geq 3} r_j^2 \right] = 0 \quad (40)$$

Clearing denominators by multiplying through by the squared characteristic polynomial $(1 + 2\eta\theta_1)^2(1 + 2\eta\theta_2)^2$ yields a polynomial equation $\Phi_{S,\hat{\beta}}(\eta) = 0$, where

$$\begin{aligned} \Phi_{S,\hat{\beta}}(\eta) = & c(1 + 2\eta\theta_1)^2(1 + 2\eta\theta_2)^2 \\ & - \eta r_1^2(1 + \theta_1\eta)(1 + 2\eta\theta_2)^2 \\ & - \eta r_2^2(1 + \theta_2\eta)(1 + 2\eta\theta_1)^2 \\ & - \eta \sum_{j \geq 3} r_j^2(1 + 2\eta\theta_1)^2(1 + 2\eta\theta_2)^2 \end{aligned} \quad (41)$$

has degree at most five.

It remains to bound the number of candidates. We proceed in three steps. First, we evaluate $\Phi_{S,\hat{\beta}}$ at λ_1 and λ_2 . We show that λ_j is a root only when $r_j = 0$ (Case 2.2), and that it is then a root of multiplicity two. Second, we count the candidates when $\Phi_{S,\hat{\beta}}$ is not the zero polynomial. A root $\eta \notin \{\lambda_1, \lambda_2\}$ gives one candidate (Case 2.3) and costs one of the five roots, while a root at λ_j gives two candidates and costs two of them by the first step. Hence the bound of five holds. Third, we treat the case where $\Phi_{S,\hat{\beta}}$ is the zero polynomial. In this case, counting roots does not apply and we solve the stationarity condition (26) directly.

[Poles of $\Phi_{S,\hat{\beta}}$] At $\eta = \lambda_1$ the factor $(1 + 2\eta\theta_1)$ vanishes, killing every term of (41) except the second one, so

$$\begin{aligned} \Phi_{S,\hat{\beta}}(\lambda_1) &= -\lambda_1 r_1^2(1 + \theta_1\lambda_1)(1 + 2\lambda_1\theta_2)^2 \\ &= -\frac{\lambda_1 r_1^2}{2} \left(1 - \frac{\theta_2}{\theta_1}\right)^2 \end{aligned} \quad (42)$$

using $\lambda_1 = -1/(2\theta_1)$, and symmetrically for λ_2 with the indices 1 and 2 exchanged. Because $\lambda_j \neq 0$ and $\theta_1\theta_2 < 0$ by (31), λ_j is a root of $\Phi_{S,\hat{\beta}}$ if and only if $r_j = 0$, which is exactly Case 2.2; and when $r_j = 0$, every remaining term of (41) carries the factor $(1 + 2\eta\theta_j)^2$, making λ_j a double root.

[Counting when $\Phi_{S,\hat{\beta}} \neq 0$] The polynomial has at most five roots counted with multiplicity. Each root $\eta \notin \{\lambda_1, \lambda_2\}$ contributes exactly one candidate through Case 2.3. Each λ_j that is a root uses

two roots and contributes at most two candidates through (35). Summing over all configurations, the number of candidates is at most five.

[Counting when $\Phi_{S,\hat{\beta}} \equiv 0$] We have $c = \Phi_{S,\hat{\beta}}(0) = 0$ and $r_1 = r_2 = 0$ by (42). The remaining (last) term of (41) forces $\sum_{j \geq 3} r_j^2 = 0$ and hence $r = 0$. The stationarity condition (26) becomes $(I + 2\eta M)u = 0$, which in the eigen coordinates reads $u_j(1 + 2\eta\theta_j) = 0$ for every j . If $\eta \notin \{\lambda_1, \lambda_2\}$, all the coefficients $1 + 2\eta\theta_j$ are nonzero and $u = 0$. If $\eta = \lambda_j$ for some $j \in \{1, 2\}$, the j -th coefficient vanishes and leaves u_j free, while every other coefficient is nonzero as shown in Case 2.2; the solution is thus $u = u_j v_j$, a multiple of the j -th eigenvector of M . Feasibility (23) then gives

$$\tilde{b}(u_j v_j, \hat{\beta}) = \theta_j u_j^2 + u_j r_j + c = \theta_j u_j^2 = 0 \quad (43)$$

using $r = 0$ and $c = 0$. $\theta_j \neq 0$ forces $u_j = 0$. So $u = 0$ in every case, and S carries the single stationary point u_0 .

In all scenarios S carries at most five stationary points, each obtained in closed form from the roots of a polynomial of degree at most five.

[Part (b)] By assumption, w^* is a global optimum of $(\mathcal{P})$ with support $S = S(w^*)$, and S is admissible by Theorem 3. Because $w_k^* > 0$ for all $k \in S$, the non-negativity constraints are inactive: w^* locally minimizes $\frac{1}{2}\|w\|^2$ subject to $\mathbf{1}^\top w = 1$ and $b(w, \hat{\beta}) = 0$ on S . Write u^* for its reduced coordinate, $w^* = u_0 + Nu^*$. We consider two scenarios:

[The constraint gradients are linearly independent] The linear independence constraint qualification holds at w^* , so u^* is a stationary point of $(\mathcal{P}_S^{(u)})$ and is enumerated by Part (a).

[The constraint gradients are linearly dependent] Dependence forces $\nabla_w b(w^*, \hat{\beta}) = \mu \mathbf{1}$ for some $\mu \in \mathbb{R}$. The chain rule gives the reduced gradient

$$\begin{aligned} \nabla_u \tilde{b}(u^*, \hat{\beta}) &= \nabla_u [b(u_0 + Nu, \hat{\beta})] \big|_{u=u^*} \\ &= N^\top \nabla_w b(w^*, \hat{\beta}) = \mu N^\top \mathbf{1} \end{aligned} \quad (44)$$

which is equal to zero since the columns of N are orthogonal to $\mathbf{1}$. Together with feasibility $\tilde{b}(u^*, \hat{\beta}) = 0$, expanding the quadratic (23) around

u^* along a direction δ with step size t gives

$$\begin{aligned} & \tilde{b}(u^* + t\delta, \hat{\beta}) \\ &= \underbrace{\tilde{b}(u^*, \hat{\beta})}_{=0} + t \underbrace{\nabla_u \tilde{b}(u^*, \hat{\beta})^\top \delta}_{=0} + \frac{t^2}{2} \delta^\top \underbrace{\nabla_u^2 \tilde{b}(u^*, \hat{\beta}) \delta}_{=2M} \\ &= t^2 \delta^\top M \delta \end{aligned} \quad (45)$$

Call δ a null direction if $\delta^\top M \delta = 0$. Along such a direction the line $u(t) = u^* + t\delta$ satisfies the constraint $\tilde{b}(\cdot, \hat{\beta}) = 0$ for every $t \in \mathbb{R}$. Because $w_k^* > 0$ for all $k \in S$, the points of this line with $|t|$ small are feasible for $(\mathcal{P})$, so $t = 0$ is a local minimizer of

$$\varphi(t) = \|u^* + t\delta\|^2 \quad (46)$$

$$= \|u^*\|^2 + 2t(u^{*\top} \delta) + t^2 \|\delta\|^2 \quad (47)$$

This function is a convex quadratic, so $\varphi'(0) = 0$, which asserts

$$u^{*\top} \delta = 0 \quad (48)$$

for every null direction δ .

If S is degenerate, then $M = 0$ and every direction is null, so (48) forces $u^* = 0$. Otherwise, M is indefinite of rank two by (31), and its null directions span the reduced space: they contain the kernel directions $\{v_j\}_{j \geq 3}$ together with $\sqrt{-\theta_2} v_1 \pm \sqrt{\theta_1} v_2$, whose span is all of $\mathbb{R}^{|S|-1}$. Orthogonality (48) to a spanning set again forces $u^* = 0$. In both cases w^* is the uniform weighting on S . Moreover $r = -2Mu^* = 0$ by (44) and $c = \tilde{b}(0, \hat{\beta}) = 0$ by feasibility, so every term of (41) vanishes and $\Phi_{S, \hat{\beta}} \equiv 0$. Part (a) enumerates $u^* = 0$ as the single stationary point on S in this situation, both when S is degenerate and when $\Phi_{S, \hat{\beta}} \equiv 0$. The global optimum is therefore contained within the candidate set enumerated in Part (a). $\square$

A.3 Proof of Theorem 3

Theorem 3 (Admissible supports). *If w_S with support S is a feasible weighting for target $\hat{\beta}$, then $\beta_{\min}(S) \leq \hat{\beta} \leq \beta_{\max}(S)$.*

Proof. For any w_S with at least two positive values, (*) gives

$$\hat{\beta} = \beta(w_S) = \frac{\sigma(a; w_S)^2}{\sigma(a; w_S)^2 + \sigma(f; w_S)^2} \quad (49)$$

Given $w_S = \{w_k\}_k$, by Proposition 1 we have

$$\hat{\beta} = \frac{\sum_{k < k'} w_k w_{k'} (a_k - a_{k'})^2}{\sum_{k < k'} w_k w_{k'} [(a_k - a_{k'})^2 + (f_k - f_{k'})^2]} \quad (50)$$

Letting

$$p_{kk'} = \frac{w_k w_{k'} [(a_k - a_{k'})^2 + (f_k - f_{k'})^2]}{\sum_{l < l'} w_l w_{l'} [(a_l - a_{l'})^2 + (f_l - f_{l'})^2]} \quad (51)$$

with $p_{kk'} \geq 0$ and $\sum_{k < k'} p_{kk'} = 1$, we substitute (12):

$$\hat{\beta} = \sum_{k < k'} p_{kk'} \beta_{kk'} \quad (52)$$

Therefore, $\hat{\beta}$ is a strict convex combination of pairwise balances over non-zero weights $\{\beta_{kk'}\}_{w_k w_{k'} > 0}$, hence strictly between their smallest and their largest, unless they all coincide, in which case it equals their common value. We have

$$\{\beta_{kk'}\}_{w_k w_{k'} > 0} \subseteq \{\beta_{kk'}\}_{kk'} \in [\beta_{\min}(S), \beta_{\max}(S)] \quad (53)$$

thus $\beta_{\min}(S) \leq \hat{\beta} \leq \beta_{\max}(S)$. $\square$

A.4 Proof of Theorem 4

Theorem 4 (Support size bound). *For any weighting w supported on S , $P_{\text{collis}}(w) \geq 1/|S|$, with equality iff w is uniform on S .*

Proof. S is the support of w , meaning $w_k = 0$ for $k \notin S$. The collision probability is defined as:

$$P_{\text{collis}}(w) = \sum_k w_k^2 = \sum_{k \in S} w_k^2 \quad (54)$$

By Cauchy–Schwarz inequality:

$$\left(\underbrace{\sum_{k \in S} w_k \cdot 1}_{=1} \right)^2 \leq \left(\underbrace{\sum_{k \in S} w_k^2}_{P_{\text{collis}}(w)} \right) \left(\underbrace{\sum_{k \in S} 1^2}_{|S|} \right) \quad (55)$$

Equality holds in the Cauchy–Schwarz inequality iff the vectors $(w_k)_{k \in S}$ and $(1)_{k \in S}$ are proportional, meaning w_k is a constant across all $k \in S$. Given $\sum_{k \in S} w_k = 1$, this forces $w_k = 1/|S|$ for all $k \in S$. Therefore, equality holds iff w is uniform on S . Dividing both sides of (55) by $|S|$ yields the claim. $\square$

A.5 Proof of Theorem 5

Theorem 5 (Exit certificate). *Let w be feasible for Problem $(\mathcal{P})$. Let $\tilde{a} = a - \mu(a)\mathbf{1}$, $\tilde{f} = f - \mu(f)\mathbf{1}$ be both nonzero, and $\rho = \frac{\tilde{a}^\top \tilde{f}}{\|\tilde{a}\| \|\tilde{f}\|}$ the adequacy–fluency Pearson correlation across systems. If there*

β -sweep range:		$\hat{\beta} \in [\beta_0 - 0.05, \beta_0 + 0.05]$				Full range $\hat{\beta} \in [\beta_{\min}, \beta_{\max}]$			
Config		# Supports	# Candidates	Success %	# FLOPs	# Supports	# Candidates	Success %	# FLOPs
Our algorithm	Theorem								
	3 4 5								
	✓ ✓ ✓	50	50	100%	1.3M	951.4K	2.1M	100%	26,304.8M
	✓ ✓ −	50	50	100%	1.3M	951.7K	2.1M	100%	26,313.2M
	− ✓ ✓	50	50	100%	1.3M	1.0M	2.3M	100%	28,146.9M
	✓ − ✓	50	50	100%	1.3M	1.8M	3.9M	100%	48,911.1M
	− ✓ −	50	50	100%	1.3M	1.0M	2.3M	100%	28,155.4M
	− − ✓	50	50	100%	1.3M	1.9M	4.3M	100%	52,520.6M
	✓ − −	3.2M	4.8M	100%	87.4B	3.0M	5.8M	100%	82,239.1M
	− − −	3.3M	4.9M	100%	88.8B	3.1M	6.2M	100%	86,307.4M
Numerical	N Tolerance								
	2 10^{-3}	—	—	18%	1.6M	—	—	16%	1.6M
	3 10^{-3}	—	—	80%	9.3M	—	—	62%	9.3M
	4 10^{-3}	—	—	100%	44.4M	—	—	88%	44.4M
	5 10^{-3}	—	—	100%	177.5M	—	—	100%	177.5M

Table 1: Ablation study of our algorithm on the ende24 dataset ($K = 16$). We sweep $\hat{\beta}$ across 50 values in each of two distinct regimes. # Supports (# Candidates) is the number of supports (candidates) visited prior to termination.

β -sweep range	Theorem		
	3	4	5
$[\beta_0 - 0.05, \beta_0 + 0.05]$	0	0	50
$[\beta_{\min}, \beta_{\max}]$	65.5K	29	19

Table 2: Number of times each of Theorems 3–5 fires over a 50-point β -sweep.

exist $\eta, \lambda \in \mathbb{R}$ such that

$$k \in S(w) : \quad w_k + \eta \frac{\partial b(w, \hat{\beta})}{\partial w_k} + \lambda = 0 \quad (\text{i})$$

$$k \notin S(w) : \quad \lambda + \eta \frac{\partial b(w, \hat{\beta})}{\partial w_k} \geq 0 \quad (\text{ii})$$

$$I_{2+2\eta} \begin{pmatrix} \hat{\beta} \|\tilde{f}\|^2 - (1-\hat{\beta})\rho^2 \|\tilde{a}\|^2, & -(1-\hat{\beta})\rho\sqrt{1-\rho^2} \|\tilde{a}\|^2 \\ -(1-\hat{\beta})\rho\sqrt{1-\rho^2} \|\tilde{a}\|^2, & -(1-\hat{\beta})(1-\rho^2) \|\tilde{a}\|^2 \end{pmatrix} \succeq 0 \quad (\text{iii})$$

where $\succeq 0$ means positive semidefinite, then w is a global optimum of Problem ($\mathcal{P}$).

Proof. The balance constraint $b(w, \hat{\beta}) = 0$ is governed by an indefinite quadratic matrix $H = \hat{\beta} f f^\top - (1 - \hat{\beta}) a a^\top$, according to (21) in the proof of Theorem 2. This makes the constraint domain non-convex. Therefore, we verify the Karush–Kuhn–Tucker (KKT; Wu, 2007) conditions, as well as the second-order sufficiency criteria adapted for non-convex quadratically constrained quadratic program (QCQPs; Park and Boyd, 2017).

[KKT Conditions] Consider the Lagrangian for Problem ($\mathcal{P}$):

$$\mathcal{L}(w, \eta, \lambda, \gamma) = \quad (\mathcal{L})$$

$$\frac{1}{2} \|w\|^2 + \eta b(w, \hat{\beta}) + \lambda \mathbf{1}^\top w = 1 - \sum_{k=1}^K \gamma_k w_k$$

where $\eta, \lambda \in \mathbb{R}$, and $\gamma_k \geq 0$ are the dual variables for the non-negativity constraints $w_k \geq 0$ so that $\gamma_k w_k = 0$ for all k (complementary slackness). Differentiating ($\mathcal{L}$) with respect to each weight w_k yields the general stationarity condition

$$\frac{\partial \mathcal{L}}{\partial w_k} = w_k + \eta \frac{\partial b(w, \hat{\beta})}{\partial w_k} + \lambda - \gamma_k = 0 \quad (56)$$

for all k . To validate this, we split the analysis into two parts based on whether $k \in S(w)$:

[Case $k \in S(w)$] By definition, $w_k > 0$. Complementary slackness forces the corresponding dual variable $\gamma_k = 0$. Substituting this into (56):

$$w_k + \eta \frac{\partial b(w, \hat{\beta})}{\partial w_k} + \lambda = 0 \quad \text{for } k \in S(w) \quad (57)$$

which is given by condition (i).

[Case $k \notin S(w)$] By substituting $w_k = 0$ into (56) and keeping only γ_k on one side:

$$\gamma_k = \lambda + \eta \frac{\partial b(w, \hat{\beta})}{\partial w_k} \geq 0 \quad \text{for } k \notin S(w) \quad (58)$$

which is given by condition (ii). Note that the inequality is given by γ_k ’s definition.

Together, (57) and (58) satisfy all KKT conditions for Problem ($\mathcal{P}$), which are necessary but not sufficient. Next, we verify second-order sufficiency criteria.

[Second-order sufficiency criteria] We have the Hessian of ($\mathcal{L}$)

$$\nabla_w^2 \mathcal{L} = I + 2\eta H \quad (59)$$

where H is the indefinite matrix $H = \hat{\beta} f f^\top - (1 - \hat{\beta}) a a^\top$, following (21). For w to be a global minimizer, the Lagrangian Hessian must be positive semi-definite along all directions v that preserve the equality constraints $\mathbf{1}^\top v = 0$ and $\nabla b(w, \hat{\beta})^\top v = 0$ (Moré and Sorensen, 1983). Through standard null-space elimination (Nocedal and Wright, 2006), this requirement projects onto the zero-sum subspace, demanding

$$v^\top (I + 2\eta H) v \geq 0 \quad (60)$$

for all vector directions v such that $\mathbf{1}^\top v = 0$.

To validate this condition without explicitly constructing high-dimensional sweeps, we invoke the structural reduction detailed in (22–23) in the proof of Theorem 2. As shown in that proof, eliminating the budget constraint via the affine transformation $w = u_0 + Nu$ restricts the quadratic form to the zero-sum subspace via the reduced matrix $M = N^\top H N$. According to Theorem 2’s proof, Case 2, M acts as the difference of two rank-1 outer products and has rank two on the relevant subspace with two non-zero eigenvalues $\theta_1 > 0 > \theta_2$. Projecting the quadratic form onto the orthonormal basis of this two-dimensional span $\{\tilde{a}, \tilde{f}\}$ transforms the test of (60) into the explicit 2×2 symmetric matrix test defined in condition (iii). That is, condition (iii) requires that $H(\eta, \hat{\beta}) \succeq 0$, and requiring $H(\eta, \hat{\beta}) \succeq 0$ is strictly equivalent to $v^\top (I + 2\eta H) v \geq 0$ for all v with $\mathbf{1}^\top v = 0$, which satisfies the second-order sufficiency criteria.

Since w satisfies KKT conditions across both active and inactive supports via (i) and (ii), and the second-order subspace sufficiency condition via (iii), w is certified to be a global optimum of Problem ($\mathcal{P}$). $\square$

B Efficiency and Ablation Analysis

We evaluate the computational efficiency of Algorithm 1 and ablate its three pruning mechanisms using the ende24 dataset¹¹ ($K = 16$). We sweep

¹¹The choice of this dataset is due to its large K . The trends and conclusions are similar across the datasets.

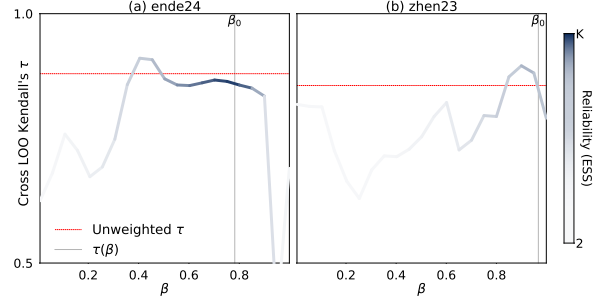


Figure 6: Leave-one-out stability with and without fixing β . The horizontal red line corresponds to the baseline values in Table 3. The blue curve shows the pooled pairwise Kendall’s τ correlation of scorer rankings after reweighting each subset to the fixed β . Curve intensity represents ESS (reliability).

$\hat{\beta}$ across 50 values in two regimes: a narrow band around the natural balance β_0 , and the full reachable range $[\beta_{\min}, \beta_{\max}]$. The former represents the practical use case, as performance is more reliable here (§4.3) and minor adjustments reflect realistic demands. The latter serves as an extreme regime to stress-test the algorithm. We compare against a numerical optimization baseline using varying grid resolutions N and a fixed tolerance of 10^{-3} . Tables 1 and 2 summarize the results.

Near natural balance. When the target $\hat{\beta}$ is close to β_0 , the optimal weighting remains close to uniform. Table 2 shows that Theorem 5 (exit certificate) triggers immediately on the full support for all 50 targets. Consequently, the search evaluates exactly one support per target. Table 1 demonstrates the resulting efficiency: solving the narrow sweep requires only 1.3M FLOPs, substantially faster than the numerical baseline required to guarantee 100% success, while strictly guaranteeing global optimality. Without early exit, the algorithm exhaustively evaluates over 3.2M supports, increasing FLOPs by nearly five orders of magnitude.

Full range. Reaching extreme $\hat{\beta}$ targets requires sparse weight vectors. In this regime, Theorem 3 (admissible supports) drives the pruning, firing over 65.5K times (Table 2) to safely skip infeasible branches. The exact algorithm demands more computation than the numerical baseline across the full range, which is the cost of guaranteeing the global minimum. The ablation confirms that combining all pruning theorems reduces the full-range computational cost by more than $3\times$ compared to the unpruned exact search.

Dataset	# Systems	Cross-LOO Kendall’s τ
ende21	13	0.8241
zhen21	13	0.6659
ted_ende	13	0.8634
ted_zhen	13	0.8774
ende22	13	0.7497
zhen22	14	0.8864
zhen23	15	0.8560
heen23	12	0.8470
ende24	16	0.8797
enes24	13	0.8815
jazh24	13	0.7931

Table 3: Pooled pairwise Kendall’s τ correlation of scorer rankings across leave-one-out (LOO) system subsets for each dataset. The sub-optimal correlation values indicate that meta-evaluation rankings are sensitive to system selection.

C Does Controlling Adequacy–Fluency Balance Stabilize Meta-Evaluation?

Meta-evaluation rankings vary across datasets. While intrinsic differences partially explain this, instability also stems directly from system selection within a single dataset. We quantify this by computing the pooled pairwise Kendall’s τ (Kendall et al., 1938) of scorer rankings across all leave-one-out (LOO) system subsets for each dataset. Table 3 reports these results. These values confirm that meta-evaluation outcomes are sensitive to the specific MT systems included in the pool.

A reasonable hypothesis is that this within-dataset instability is partly driven by varying adequacy–fluency balances across subsets. To test this, we evaluate whether fixing β stabilizes the scorer rankings across the LOO subsets.¹² Our results reject this hypothesis. As illustrated for two representative datasets in Figure 6, precisely controlling β does not resolve the ranking variance. While the adequacy–fluency balance may contribute to the instability of specific datasets, it is not the primary driver of the system-selection sensitivity observed in recent WMT evaluations. Identifying the root cause of this instability remains an open problem beyond the scope of this work.

D Dataset Statistics

Tables 4–7 report descriptive statistics for the eleven datasets used throughout this work. Ta-

ble 4 covers pool size and the balance parameter. Table 5 summarizes the distribution of pairwise balances $\beta_{kk'}$. Table 6 reports system-level Adequacy MQM and Fluency MQM statistics, together with their correlation and covariance across systems. Table 7 lists by name the base scorers augmented for each dataset.

E Main Results for All Datasets

Figure 7 presents the main results across all datasets, reproducing the figures from the main text for completeness. The observed trends remain consistent throughout, demonstrating our method’s generalizability.

F Main Results for Pearson Correlation

Figure 8 mirrors the experiments from §4.3, using Pearson correlation as the meta-metric rather than SPA. Markers representing Shayegh et al. (2025) are omitted, as Pearson correlation falls outside the scope of their work. These findings align closely with our SPA evaluation, confirming the robustness of our approach.

G Variance Analysis for All Datasets

Figure 9 presents the variance analysis heatmaps across all datasets, reproducing the figures from the main text for completeness. The observed trends remain consistent throughout, demonstrating the reliability of β as a control parameter.

¹²In preliminary studies, we also controlled the adequacy-over-fluency preference directly (see §4) to account for different subsets possessing different balance– β relations. This alternative control also failed to resolve the instability.

Dataset	#Systems	#Scorers	#Segments	β_0	$\beta_{\min}$	$\beta_{\max}$	$\beta_{\max} - \beta_{\min}$
ende24	16	30	487	0.7821	0.0074	1.0000	0.9926
enes24	13	30	623	0.9657	0.0005	1.0000	0.9994
jazh24	13	31	560	0.9954	0.4059	1.0000	0.5941
heen23	12	43	820	0.9868	0.0022	1.0000	0.9978
zhen23	15	43	1177	0.9672	0.0034	1.0000	0.9966
ende22	13	31	1315	0.9338	0.0010	1.0000	0.9990
zhen22	14	31	1875	0.9767	0.0071	1.0000	0.9928
ende21	13	27	527	0.2273	0.0000	0.9996	0.9996
zhen21	13	28	650	0.7054	0.0008	1.0000	0.9991
ted_ende	13	27	529	0.3802	0.0034	1.0000	0.9966
ted_zhen	13	28	529	0.4921	0.0063	0.9995	0.9932

Table 4: Dataset size and reachable balance range. #Scorers is the number of base scorers used for augmentation, listed by name in Table 7.

Dataset	#Pairs	Mean	Median	Std	IQR
ende24	120	0.6223	0.6883	0.3083	0.5019
enes24	78	0.8444	0.9705	0.2642	0.1335
jazh24	78	0.9317	0.9922	0.1538	0.0292
heen23	66	0.8971	0.9919	0.2065	0.0917
zhen23	105	0.8392	0.9508	0.2416	0.1835
ende22	78	0.8049	0.9115	0.2523	0.2376
zhen22	91	0.8576	0.9499	0.2160	0.1371
ende21	78	0.3543	0.3105	0.2916	0.4552
zhen21	78	0.5984	0.7165	0.3347	0.4885
ted_ende	78	0.4672	0.3872	0.3667	0.7545
ted_zhen	78	0.5278	0.5340	0.3273	0.5987

Table 5: Statistics describing the distribution of the pairwise balance $\beta_{kk'}$ across system pairs. IQR: their interquartile range, i.e. the 75th percentile minus the 25th.

Dataset	$\mu(a)$	$\sigma(a)^2$	$\mu(f)$	$\sigma(f)^2$	ρ	$\Sigma(a, f)$
ende24	-1.3745	0.4475	-1.1275	0.1247	0.7210	0.1703
enes24	-0.4547	0.1167	-0.1262	0.0042	0.5531	0.0122
jazh24	-2.5512	2.3894	-0.3477	0.0110	0.5749	0.0933
heen23	-2.0684	0.7275	-0.3771	0.0097	0.4214	0.0355
zhen23	-2.7212	1.5735	-1.4967	0.0534	0.2707	0.0784
ende22	-0.8175	0.0494	-0.4051	0.0035	0.6182	0.0081
zhen22	-2.6217	0.8798	-0.6556	0.0210	0.8144	0.1108
ende21	-0.5341	0.0364	-1.1930	0.1238	0.6377	0.0428
zhen21	-3.3916	0.1510	-1.7004	0.0631	0.0837	0.0082
ted_ende	-0.7215	0.0262	-0.8389	0.0426	0.2913	0.0097
ted_zhen	-1.1248	0.0461	-1.0927	0.0476	0.7116	0.0333

Table 6: System-level Adequacy and Fluency MQM. Every statistic is taken over the K systems of a single dataset, at the uniform weighting $w = \frac{1}{K}\mathbf{1}$, on the system-level score vectors a (Adequacy MQM) and f (Fluency MQM). ρ is the Pearson correlation between a and f across systems. $\Sigma(a, f)$: the covariance between a and f across those same systems. Means are negative because MQM is defined as a negated error weight.

Dataset	Base scorers
ende24	BERTScore, BLCOM, BLCOM_1, BLEU, BLEURT-20, COMET-22, CometKiwi, CometKiwi-XXL, MEE4, MetricX-24, MetricX-24-Hybrid, MetricX-24-Hybrid-QE, MetricX-24-QE, PrismRefMedium, PrismRefSmall, XCOMET, XCOMET-QE, XLSimDA, YiSi-1, bright-qe, chrF, chrFS, damonmonli, gemba_esa, metametrics_mt_mqm_hybrid_kendall, metametrics_mt_mqm_kendall, metametrics_mt_mqm_qe_kendall.seg.s, monmonli, sentinel-cand-mqm, spBLEU
enes24	BERTScore, BLCOM, BLCOM_1, BLEU, BLEURT-20, COMET-22, CometKiwi, CometKiwi-XXL, MEE4, MetricX-24, MetricX-24-Hybrid, MetricX-24-Hybrid-QE, MetricX-24-QE, PrismRefMedium, PrismRefSmall, XCOMET, XCOMET-QE, XLSimDA, YiSi-1, bright-qe, chrF, chrFS, damonmonli, gemba_esa, metametrics_mt_mqm_hybrid_kendall, metametrics_mt_mqm_kendall, metametrics_mt_mqm_qe_kendall.seg.s, monmonli, sentinel-cand-mqm, spBLEU
jazh24	BERTScore, BLCOM, BLCOM_1, BLEU, BLEURT-20, COMET-22, CometKiwi, CometKiwi-XXL, MEE4, MetricX-24, MetricX-24-Hybrid, MetricX-24-Hybrid-QE, MetricX-24-QE, PrismRefMedium, PrismRefSmall, XCOMET, XCOMET-QE, XLSimDA, YiSi-1, bright-qe, chrF, chrFS, damonmonli, gemba_esa, metametrics_mt_mqm_hybrid_kendall, metametrics_mt_mqm_qe_kendall.seg.s, metametrics_mt_mqm_qe_same_source_t, metametrics_mt_mqm_same_source_targ, monmonli, sentinel-cand-mqm, spBLEU
heen23	BERTscore, BLEU, BLEURT-20, COMET, Calibri-COMET22, Calibri-COMET22-QE, CometKiwi, CometKiwi-XL, CometKiwi-XXL, GEMBA-MQM, KG-BERTScore, MEE4, MS-COMET-QE-22, MaTESe, MetricX-23, MetricX-23-QE, MetricX-23-QE-b, MetricX-23-QE-c, MetricX-23-b, MetricX-23-c, Random-sysname, XCOMET-Ensemble, XCOMET-QE-Ensemble, XCOMET-XL, XCOMET-XXL, XLSim, YiSi-1, chrF, cometoid22-wmt21, cometoid22-wmt22, cometoid22-wmt23, docWMT22CometDA, docWMT22CometKiwiDA, eBLEU, embed_llama, f200spBLEU, instructscore, mbr-metricx-qe, mre-score-labse-regular, prismRef, prismSrc, sescoreX, tokengram_F
zhen23	BERTscore, BLEU, BLEURT-20, COMET, Calibri-COMET22, Calibri-COMET22-QE, CometKiwi, CometKiwi-XL, CometKiwi-XXL, GEMBA-MQM, KG-BERTScore, MEE4, MS-COMET-QE-22, MaTESe, MetricX-23, MetricX-23-QE, MetricX-23-QE-b, MetricX-23-QE-c, MetricX-23-b, MetricX-23-c, Random-sysname, XCOMET-Ensemble, XCOMET-QE-Ensemble, XCOMET-XL, XCOMET-XXL, XLSim, YiSi-1, chrF, cometoid22-wmt21, cometoid22-wmt22, cometoid22-wmt23, docWMT22CometDA, docWMT22CometKiwiDA, eBLEU, embed_llama, f200spBLEU, instructscore, mbr-metricx-qe, mre-score-labse-regular, prismRef, prismSrc, sescoreX, tokengram_F
ende22	BERTScore, BLEU, BLEURT-20, COMET-20, COMET-22, COMET-QE, COMETKiwi, Cross-QE, HWTSC-TLM, HWTSC-Teacher-Sim, KG-BERTScore, MATESE, MATESE-QE, MEE, MEE2, MEE4, MS-COMET-22, MS-COMET-QE-22, REUSE, SEScore, UniTE, UniTE-ref, UniTE-src, YiSi-1, chrF, f101spBLEU, f200spBLEU, metricx_xl_DA_2019, metricx_xl_MQM_2020, metricx_xxl_DA_2019, metricx_xxl_MQM_2020
zhen22	BERTScore, BLEU, BLEURT-20, COMET-20, COMET-22, COMET-QE, COMETKiwi, Cross-QE, HWTSC-TLM, HWTSC-Teacher-Sim, KG-BERTScore, MATESE, MATESE-QE, MEE, MEE2, MEE4, MS-COMET-22, MS-COMET-QE-22, REUSE, SEScore, UniTE, UniTE-ref, UniTE-src, YiSi-1, chrF, f101spBLEU, f200spBLEU, metricx_xl_DA_2019, metricx_xl_MQM_2020, metricx_xxl_DA_2019, metricx_xxl_MQM_2020
ende21	BERTScore, C-SPEC, C-SPECpn, COMET-DA_2020, COMET-DA_2021, COMET-MQM_2021, COMET-QE-DA_2021, COMET-QE-MQM_2021, COMETinho-DA, COMETinho-MQM, MEE, MEE2, MTEQA, OpenKiwi-MQM, Prism, TER, YiSi-1, YiSi-2, bleurt-20, bleurt-21-beta, chrF, cushLEPOR(LM), cushLEPOR(pSQM), hLEPOR, regEMT, regEMT-baseline, sentBLEU
zhen21	BERTScore, C-SPEC, C-SPECpn, COMET-DA_2020, COMET-DA_2021, COMET-MQM_2021, COMET-QE-DA_2021, COMET-QE-MQM_2021, COMETinho-DA, COMETinho-MQM, MEE, MEE2, MTEQA, OpenKiwi-MQM, Prism, RoBLEURT, TER, YiSi-1, YiSi-2, bleurt-20, bleurt-21-beta, chrF, cushLEPOR(LM), cushLEPOR(pSQM), hLEPOR, regEMT, regEMT-baseline, sentBLEU
ted_ende	BERTScore, C-SPEC, C-SPECpn, COMET-DA_2020, COMET-DA_2021, COMET-MQM_2021, COMET-QE-DA_2021, COMET-QE-MQM_2021, COMETinho-DA, COMETinho-MQM, MEE, MEE2, MTEQA, OpenKiwi-MQM, Prism, TER, YiSi-1, YiSi-2, bleurt-20, bleurt-21-beta, chrF, cushLEPOR(LM), cushLEPOR(pSQM), hLEPOR, regEMT, regEMT-baseline, sentBLEU
ted_zhen	BERTScore, C-SPEC, C-SPECpn, COMET-DA_2020, COMET-DA_2021, COMET-MQM_2021, COMET-QE-DA_2021, COMET-QE-MQM_2021, COMETinho-DA, COMETinho-MQM, MEE, MEE2, MTEQA, OpenKiwi-MQM, Prism, RoBLEURT, TER, YiSi-1, YiSi-2, bleurt-20, bleurt-21-beta, chrF, cushLEPOR(LM), cushLEPOR(pSQM), hLEPOR, regEMT, regEMT-baseline, sentBLEU

Table 7: The base scorers augmented for each dataset, by name. This follows the selection described in §4. Names are reproduced exactly as released by WMT.

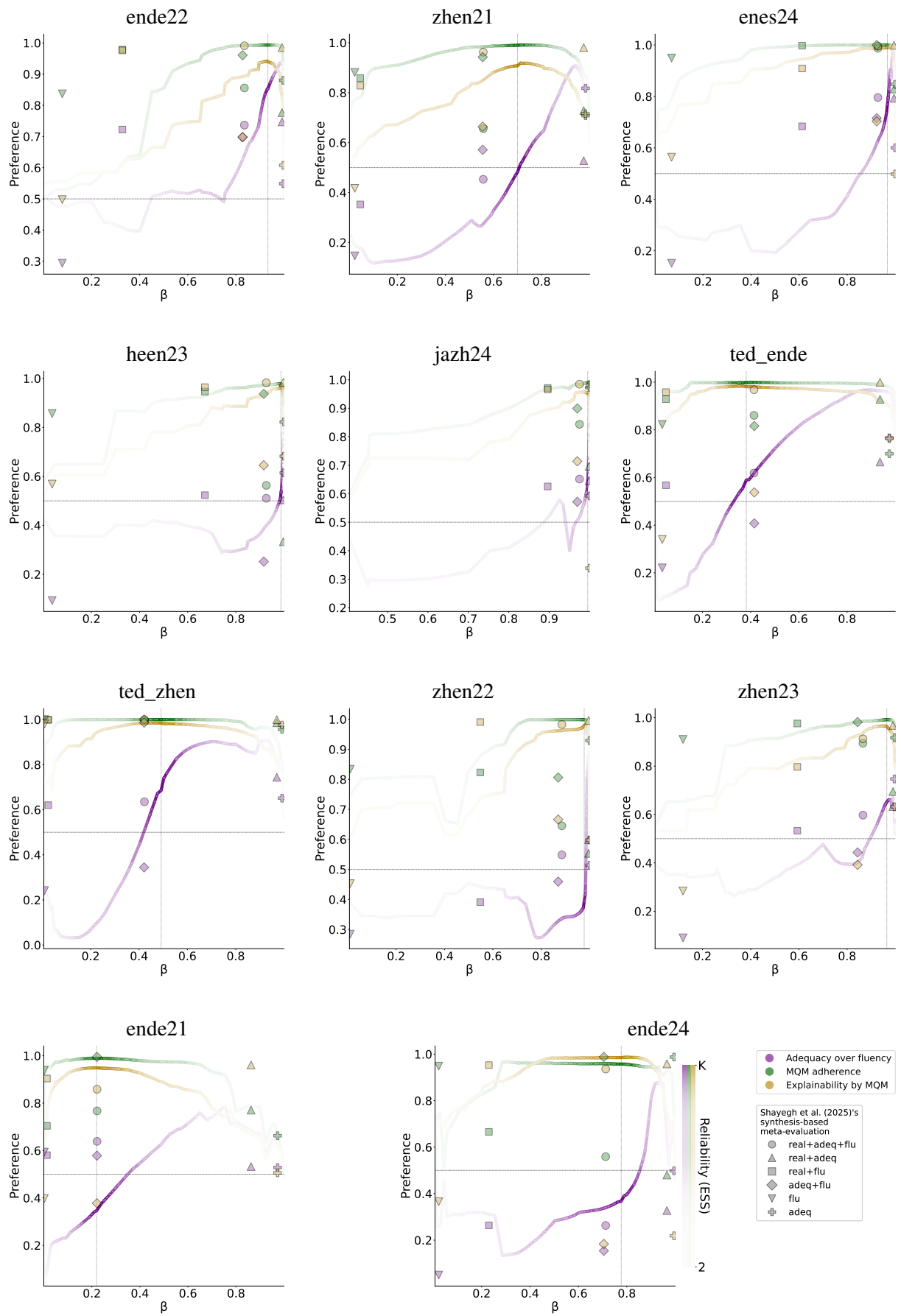


Figure 7: Main results on all datasets. Layout follows Figure 3.

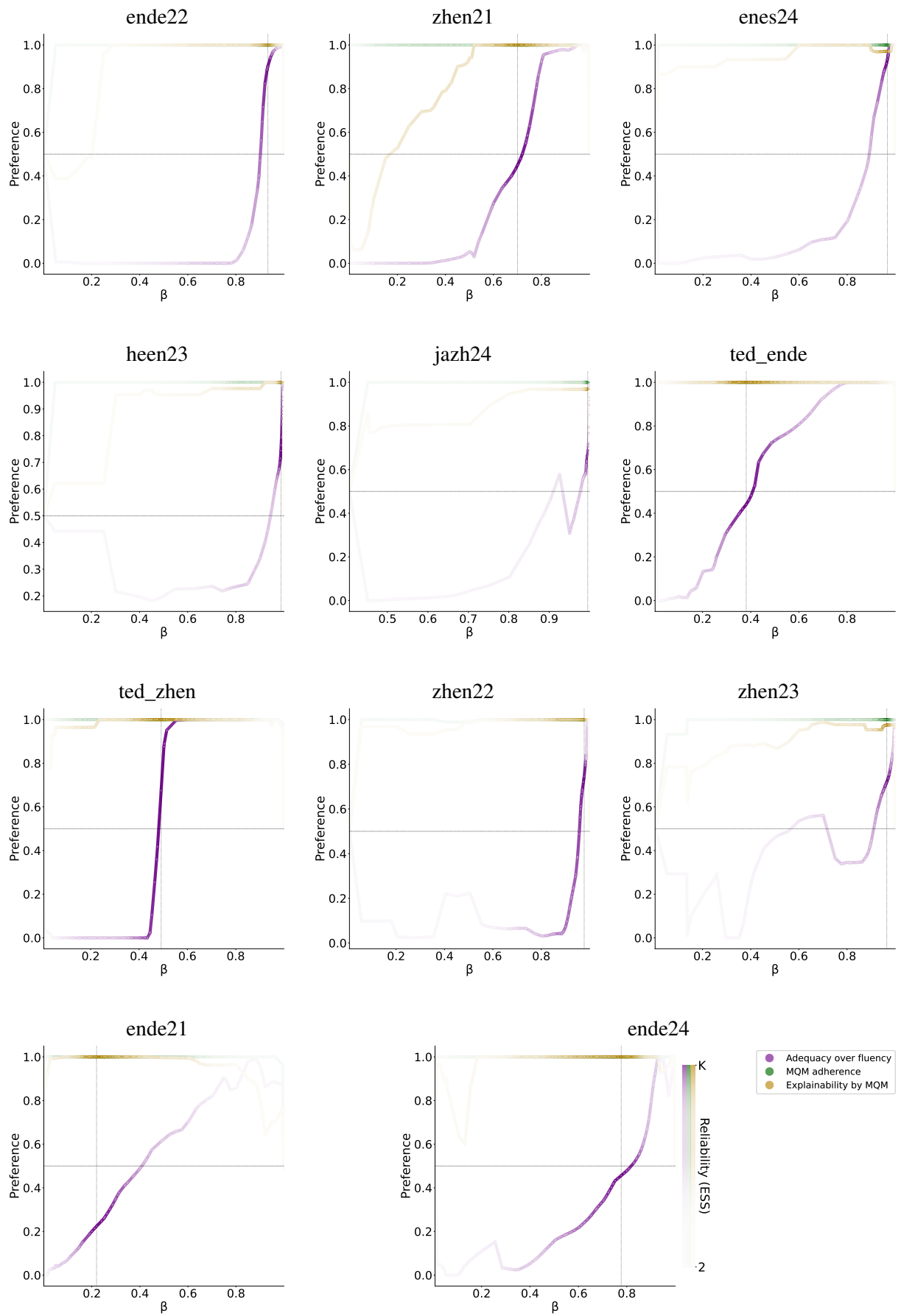


Figure 8: Main results with Pearson as the meta-metric. Layout follows Figure 3.

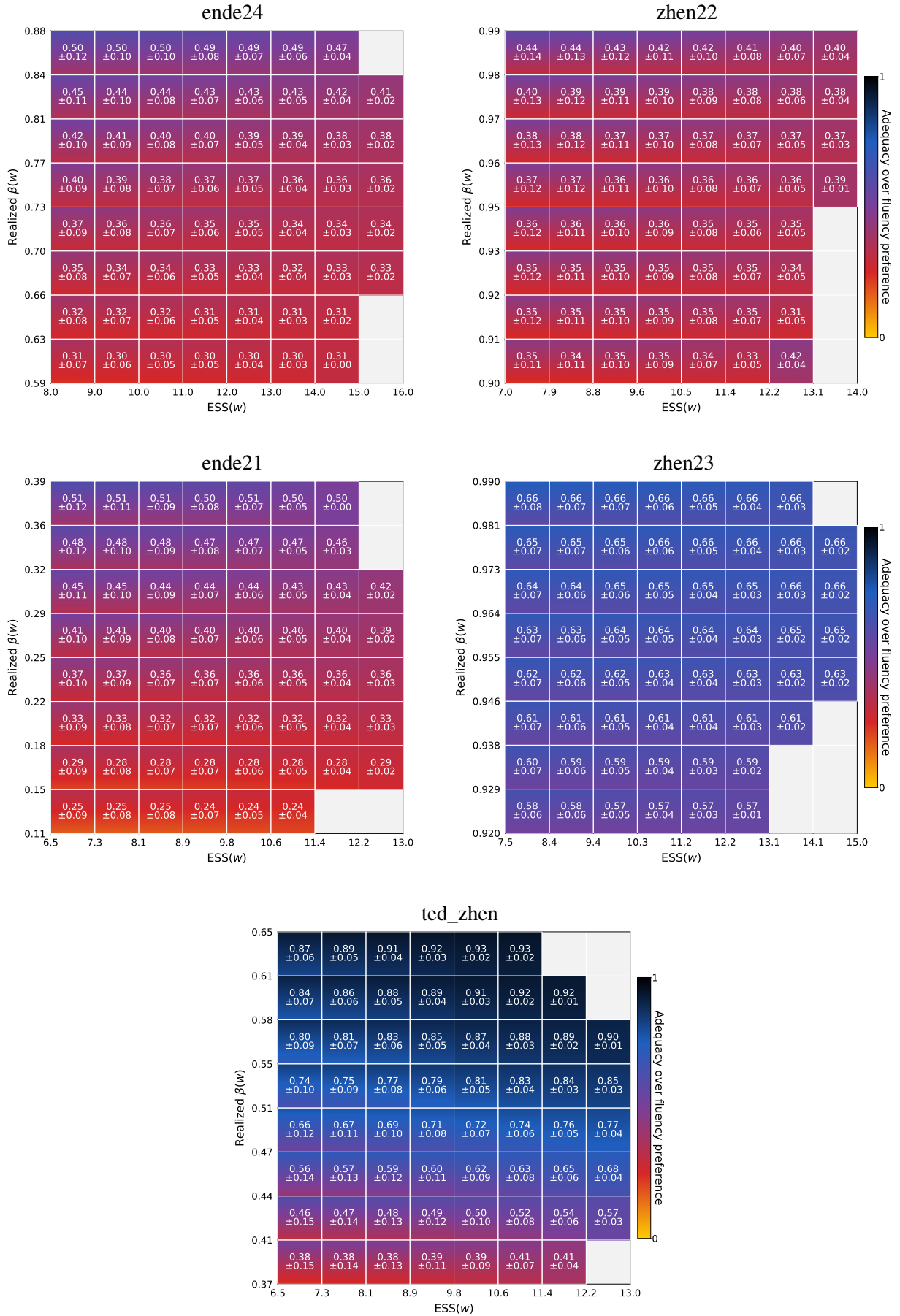


Figure 9: Variance analysis results for all datasets. Layout follows Figure 4.