Which Metrics Save the Most Human Annotation? Prediction-Powered Evaluation and Meta-Evaluation

Anonymous ACL submission

Abstract

Across various non-verifiable tasks, human evaluation is reliable but expensive, while automatic metrics are more scalable but often biased. We propose *prediction-powered evaluation*, a framework that combines limited human judgments with large-scale automatic scores to obtain data-efficient system comparisons that are provably unbiased. Specifically, we develop parametric and non-parametric prediction-powered procedures, analyze the efficiency trade-off between paired and unpaired designs, and empirically validate the framework on six WMT datasets. We further introduce the *Prediction-Powered Saving Ratio (PPSR)*, a meta-metric that measures how much human annotation an automatic metric can save when used within prediction-powered evaluation. PPSR directly targets metric utility for human-efficient evaluation and yields more discriminative and stable metric rankings than existing system-level meta-metrics. Overall, our new paradigm reframes automatic metrics as tools for reducing human annotation cost rather than replacing human judgment, and applies broadly to non-verifiable tasks where automatic metrics are available.

1 Introduction

System comparison is a central question in AI evaluation: we often want to determine which model or system performs better, and by how much. For non-verifiable tasks such as machine translation (MT), human evaluation is typically regarded as the gold standard. However, human evaluation is expensive, resulting in limited sample sizes, which often leave system comparisons statistically underpowered (Card et al., 2020; Howcroft and Rieser, 2021). In contrast, automatic metrics, including LLM-based judges, can be applied at large scale and low cost (Wei and Jia, 2021). However, automatic metrics do not always match gold standard human evaluations. These metrics are unreliable to use alone, but it is wasteful to ignore them entirely.

We argue that this dilemma arises largely because automatic evaluation has traditionally been framed as a proxy for human evaluation (Mathur et al., 2020a). In practice, this framing leads to either *human-only* evaluation or *auto-only* evaluation, with relatively little use of approaches that combine the two, such as using human judgments to debias automatic evaluation (Gao et al., 2024). Prediction-powered inference (PPI, Angelopoulos et al., 2023a,b; Ji et al., 2025; Chen et al., 2026a), a line of work from the statistical literature, offers a formal way to combine human and automatic evaluation. Specifically, it shifts the goal of automatic evaluation from *human approximation* to *variance reduction*. This increases statistical power and lowers the amount of human annotation needed for accurate system comparison.

However, applying PPI to MT system comparison is not straightforward. We identify three design choices that have long been central to the MT community. The first is *parametric vs. non-parametric inference*: standard PPI yields parametric hypothesis tests and confidence intervals, whereas WMT shared tasks have traditionally relied on non-parametric procedures, such as the sign test (Callison-Burch et al., 2012), bootstrap resampling (Bojar et al., 2016), and the Wilcoxon signed-rank test (Kocmi et al., 2025). The second is *paired vs. unpaired designs*: standard PPI naturally leads to a paired design, while WMT shared tasks use both paired designs (Kocmi et al., 2024, 2025) and unpaired designs (Barrault et al., 2020; Akhbardeh et al., 2021; Kocmi et al., 2022, 2023) depending on year. The last is *meta-evaluation*—given many automatic metrics to choose from, which should be used? While existing meta-metrics, such as SPA (Thompson et al., 2024), answer this question by minimizing disagreement with human-evaluation, these are ill-formed for PPI because PPI naturally produces unbiased estimates.

To address these design choices, we make the following contributions. **(1)** We apply PPI to system comparison with pointwise automatic metrics (§4.1), and we validate its effectiveness on six WMT datasets (§6.2). Prediction-powered evaluation is unbiased, making it more reliable than automatic-only evaluation, and yet more data-efficient than human-only evaluation. **(2)** Motivated by traditional WMT evaluation, we propose a non-parametric hypothesis test (§4.2) and empirically validate it (§6.3). This handles the case where the asymptotic normality required by standard PPI is not well satisfied. **(3)** We provide a theoretical comparison of the data efficiency of paired and unpaired designs used in WMT evaluation (§6.4). We identify key factors driving evaluation differences and empirically justify the choice of a paired design. **(4)** Building on our data-efficiency analysis, we propose the **Prediction-Powered Saving Ratio** (PPSR), a new meta-metric for prediction-powered evaluation (§5). This meta-metric reframes traditional meta-evaluation in terms of the amount of human annotation saved by a given automatic metric, rather than focusing how well automatic metrics align with human annotations. PPSR exhibits better discriminative power and ranking stability than existing meta-metrics (§6.5 and §6.7).

More broadly, prediction-powered evaluation reframes the long-standing trade-off between cheap, biased automatic evaluation and expensive, gold-standard human evaluation as a smooth annotation-savings curve, with PPSR providing the principled instrument to read off where any given automatic metric sits on that curve. Notably, the framework is *not* specific to MT: it applies wherever automatic metrics are available for non-verifiable tasks.

2 Related Work

PPI for evaluation. Existing PPI-based evaluation work divides by whether the automatic metric is *pairwise*, targeting statistics such as win rate and Bradley–Terry scores (Chatzi et al., 2024; Boyeau et al., 2025; Zhou et al., 2025), or *pointwise*, where prior work (Chaganty et al., 2018; Saad-Falcon et al., 2024; Fisch et al., 2024; Dorner et al., 2025; Guerdan et al., 2025; Chen et al., 2026b) estimates the quality of a single model (a closely related line uses Rogan–Gladen (Rogan and Gladen, 1978) for the same target (Lee et al., 2025; Feng et al., 2026)). We focus on the pointwise case, the dominant class in MT/NLG, but our target is the score *difference* between systems, which is what MT system comparison requires and which the above lines do not directly address. We additionally study two design axes crucial to MT evaluation: parametric vs. non-parametric and paired vs. unpaired.

Hypothesis tests in NLP evaluation. Dror et al. (2018) formalized algorithm comparison in NLP as hypothesis testing. The dominant parametric choice for MT/NLG is the paired t-test (van der Lee et al., 2021); non-parametric alternatives including Wilcoxon signed-rank (Kocmi et al., 2025) and rank-sum (Kocmi et al., 2023), paired bootstrap (Koehn, 2004), and the paired (approximate randomization) permutation test (Graham et al., 2014), are widely used and often preferred. All these tests are designed for human-only or auto-only settings. The prediction-powered tests we study cover the combined setting and both parametric and non-parametric variants.

Meta-metrics in MT and NLG. A wide range of meta-metrics has been proposed for evaluating automatic metrics, reflecting the many possible combinations of grouping methods and agreement functions (Gao et al., 2025; Deutsch et al., 2023; Perrella et al., 2024; DiIanni and Deutsch, 2025). For system comparison, the common system-level choices, including Pearson's r (Ma et al., 2019), Spearman's ρ (Machácek and Bojar, 2013), Kendall's τ (Mathur et al., 2020b), pairwise accuracy (Kocmi et al., 2021), and soft pairwise accuracy (SPA) (Thompson et al., 2024), all directly measure agreement with human judgments. PPSR is fundamentally different: derived from PPI, it quantifies the proportion of human judgments an automatic metric can save in system comparison while preserving the same statistical power. Quantities analogous to PPSR appear in Chaganty et al. (2018) and Zhou et al. (2025) from a control-variate perspective, but are not formalized as meta-metrics.

3 Problem and Background

3.1 System Comparison

Suppose there are N systems to be compared, denoted by $\{S_i\}_{i=1}^N$, and M source inputs (i.e. segments), denoted by $\{D_j\}_{j=1}^M$. The output of the i-th system on the j-th input is denoted by $X_{ij} = S_i(D_j)$. For all the system outputs on L of the inputs, we conduct human evaluation and obtain their human scores, denoted by $\{Y_{ij}\}_{i=1,j=1}^{N,L}$. The human scores of the remaining $U = M - L$

outputs of each system are not available. In contrast, we obtain metric scores for all system outputs using a pointwise automatic metric F, denoted by $\{F(X_{ij})\}_{i=1,j=1}^{N,M}$[1].

For any pair of systems, such as S_1 and S_2, our goal is to estimate their true human score difference $\delta_{1,2} = \mathbb{E}[Y_1 - Y_2]$ (Wei and Jia, 2021). **Human-only evaluation** uses only human scores:

$$\widehat{\delta_{1,2}^{H}} = \frac{1}{L}\sum\nolimits_{j=1}^{L}(Y_{1j} - Y_{2j}). \quad (1)$$

If $\{Y_{1j}\}_{j=1}^{L}$ and $\{Y_{2j}\}_{j=1}^{L}$ are regarded as independent and identically distributed random variables separately[2], it's easy to know $\widehat{\delta_{1,2}^{H}}$ is an unbiased estimator with $\mathrm{Var}[\widehat{\delta_{1,2}^{H}}] = \frac{1}{L}\mathrm{Var}[Y_1 - Y_2]$.

Auto-only evaluation uses only metric scores:

$$\widehat{\delta_{1,2}^{A}} = \frac{1}{M}\sum\nolimits_{j=1}^{M}(F(X_{1j}) - F(X_{2j})). \quad (2)$$

This estimator is **biased** because the automatic metric F is generally biased. Its variance $\mathrm{Var}[\widehat{\delta_{1,2}^{A}}] = \frac{1}{M}\mathrm{Var}[F(X_1) - F(X_2)]$ is always small because M is generally very large.

3.2 Meta-Evaluation: Metric Comparison

The comparisons above concern MT systems. We now turn to meta-evaluation, where the objects of comparison are automatic metrics. Suppose that we wish to compare K automatic metrics, denoted by $\{F_k\}_{k=1}^{K}$. For each metric F_k, a meta-metric C evaluates its agreement with human judgments using the metric scores $\{F_k(X_{ij})\}_{i=1,j=1}^{N,L}$ and the corresponding human scores $\{Y_{ij}\}_{i=1,j=1}^{N,L}$. Formally, the meta-evaluation score of F_k can be written as $C\left(\left\{\left(Y_{ij}, F_k(X_{ij})\right)\right\}_{i=1,j=1}^{N,L}\right)$. Typically, C returns a value in $[0, 1]$, with larger values indicating that F_k aligns more closely with human judgments according to the criterion encoded by C. These meta-evaluation scores can therefore be used both to compare pairs of automatic metrics and to induce an overall ranking among them. Criteria for assessing a meta-metric include the interpretability of its meaning (Perrella et al., 2024), its discriminative power across different automatic metrics, and the stability of the resulting metric rankings (Thompson et al., 2024).

[1]For convenience, we omit possible additional inputs to F other than X, such as D.

[2]Here Y_1 and Y_2 are random variables with the same distribution as $\{Y_{1j}\}_{j=1}^{L}$ and $\{Y_{2j}\}_{j=1}^{L}$ separately. The same applies to $F(X_1)$ and $F(X_2)$.

3.3 Prediction-Powered Inference

Prediction-powered inference (PPI) combines a small labeled sample with a larger unlabeled sample equipped with predictions from an arbitrary predictive model (Angelopoulos et al., 2023a). Consider a generic estimation problem with target $\theta = \mathbb{E}[Z]$, where Z is expensive to observe. Let W be cheaply observed information and let $g(W)$ be a prediction of Z. Suppose we observe L labeled examples $\{(W_j, Z_j)\}_{j=1}^{L}$ and U additional unlabeled examples $\{W_j\}_{j=L+1}^{L+U}$, for which only $g(W_j)$ is available. A classical estimator of θ that only uses labeled examples is $\widehat{\theta}^{H} = \frac{1}{L}\sum_{j=1}^{L} Z_j$. In constrast, a prediction-powered estimator of θ is

$$\widehat{\theta}^{PP} = \frac{\lambda}{U}\sum_{j=L+1}^{L+U} g(W_j) + \frac{1}{L}\sum_{j=1}^{L}\left(Z_j - \lambda g(W_j)\right), \quad (3)$$

where λ is a tunable parameter. The first term uses abundant predictions, while the second term corrects their bias using labeled observations. $\widehat{\theta}^{PP}$ remains unbiased for θ for any fixed λ, even when g is biased. An appropriate choice of λ can guarantee that $\mathrm{Var}[\widehat{\theta^{PP}}] \leq \mathrm{Var}[\widehat{\theta^{H}}]$ (Angelopoulos et al., 2023b), which means shorter confidence intervals and hypothesis tests with higher power.

4 Prediction-Powered Evaluation for MT

We now instantiate the generic PPI estimator of §3.3 for paired MT system comparison and derive the associated parametric confidence intervals and hypothesis tests (§4.1). Two extensions follow: a paired non-parametric hypothesis test for the small-L regime where asymptotic normality is unreliable (§4.2), and an ablation contrasting the paired default with its unpaired counterpart (§4.3).

4.1 Paired Parametric Design (Default)

Map the generic PPI setup of §3.3 to system comparison by setting $Z_j = Y_{1j} - Y_{2j}$ (the expensive per-segment human-score difference) and $g(W_j) = F(X_{1j}) - F(X_{2j})$ (the cheap metric-score difference). The target $\theta = \mathbb{E}[Z] = \delta_{1,2}$ is exactly the between-system effect of interest. Substituting into (3) yields the **prediction-powered paired estimator**

$$\begin{aligned}\widehat{\delta_{1,2}^{PP}} = &\frac{\lambda}{U}\sum_{j=L+1}^{L+U}\left(F(X_{1j}) - F(X_{2j})\right) \\ &+ \frac{1}{L}\sum_{j=1}^{L}\Big[(Y_{1j} - Y_{2j}) - \lambda\big(F(X_{1j}) - F(X_{2j})\big)\Big].\end{aligned} \quad (4)$$

At $\lambda = 0$, (4) reduces to the human-only paired estimator $\widehat{\delta^H_{1,2}}$; at $\lambda = 1$, it equals the auto-only estimator $\widehat{\delta^A_{1,2}}$ plus a rectifier term estimated on the labeled set. For *any* $\lambda \in \mathbb{R}$, $\widehat{\delta^{PP}_{1,2}}$ is **unbiased** for $\delta_{1,2}$ because the two λ-scaled metric terms have equal expectation and cancel; this holds regardless of the bias of F (Appendix A provides the proof).

Variance and optimal λ. Following Angelopoulos et al. (2023b); Chen et al. (2026a), the variance-minimizing tuning parameter $\lambda^\star$ has a closed form in terms of $\mathrm{Cov}[Y_1 - Y_2, F(X_1) - F(X_2)]$ and $\mathrm{Var}[F(X_1) - F(X_2)]$, yielding

$$\mathrm{Var}\left[\widehat{\delta^{PP}_{1,2}}\right] = \frac{1}{L}\mathrm{Var}[Y_1 - Y_2] - \frac{U\big(\mathrm{Cov}[Y_1 - Y_2,\, F(X_1) - F(X_2)]\big)^2}{L(L+U)\,\mathrm{Var}[F(X_1) - F(X_2)]}. \tag{5}$$

The second term in (5) is the variance reduction PPI buys over the human-only paired estimator. It is non-negative, vanishes when the metric-score difference is uncorrelated with the human-score difference. Thus we have $\mathrm{Var}[\widehat{\delta^{PP}_{1,2}}] \le \mathrm{Var}[\widehat{\delta^H_{1,2}}]$, so PPI is never worse than human-only evaluation under the optimal $\lambda^\star$, even when F is highly biased or weakly informative.

Confidence intervals and hypothesis test. A central-limit-theorem argument gives the $100(1-\alpha)\%$ CI for human-only and prediction-powered cases (Appendix B provides the proof)[3]:

$$\widehat{\delta^g_{1,2}} \pm z_{\alpha/2}\sqrt{\widehat{\mathrm{Var}}[\widehat{\delta^g_{1,2}}]}, \quad g = \{H, PP\},$$

where $\widehat{\mathrm{Var}}[\cdot]$ plugs sample variance. The corresponding parametric paired Z-test are summarized as Alg. 1 and Alg. 3 in the Appendix.

4.2 Paired Non-Parametric Test

The parametric CI and Z-test of §4.1 inherits the standard PPI asymptotic guarantee only when both the terms in (4) are well approximated by a normal distribution. In MT system comparison this is not always the case: human scores are discrete, and the sample size is small, etc. WMT shared tasks have for two decades preferred non-parametric procedures possibly for these reasons, but the existing prediction-powered literature has no off-the-shelf non-parametric test for paired system comparison[4].

To fill this gap, We extend the classical paired permutation test (Alg. 4 and Alg. 5 in the Appendix) (Graham et al., 2014) to the prediction-powered setting (Alg. 6 in the appendix). The test is finite-sample valid: under H_0, the permutation distribution exactly characterizes the null, so the asymptotic-normality assumption of §4.1 is not required.

4.3 Unpaired Design

Sections 4.1 and 4.2 assume a paired design: both systems are evaluated on the same source inputs, and $\widehat{\delta^{PP}_{1,2}}$ is built from per-segment differences. WMT17–WMT23 instead used an unpaired design at the test level (Wilcoxon rank-sum), and the unpaired design is the natural fallback whenever the two systems are evaluated on disjoint inputs. We treat the paired-vs-unpaired choice as an ablation of a design parameter and quantify the cost of the unpaired option, both for human-only and prediction-powered evaluation.

Unpaired estimators. Let I^L_i, I^U_i index the labeled and unlabeled inputs for system S_i (the paired design corresponds to $I^L_1 = I^L_2, I^U_1 = I^U_2$). The unpaired human-only estimator,

$$\widehat{\delta^{H,\mathrm{Un}}_{1,2}} = \frac{1}{|I^L_1|}\sum_{j\in I^L_1} Y_{1j} - \frac{1}{|I^L_2|}\sum_{j\in I^L_2} Y_{2j},$$

treats the two systems' scores as independent. The unpaired prediction-powered estimator instantiates the generic PPI of (3) separately for each system, with two tuning parameters λ_1, λ_2:

$$\begin{aligned}\widehat{\delta^{PP,\mathrm{Un}}_{1,2}} = {} & \frac{\lambda_1}{|I^U_1|}\sum_{j\in I^U_1} F(X_{1j}) + \frac{1}{|I^L_1|}\sum_{j\in I^L_1}\big(Y_{1j} - \lambda_1 F(X_{1j})\big) \\ & - \frac{\lambda_2}{|I^U_2|}\sum_{j\in I^U_2} F(X_{2j}) - \frac{1}{|I^L_2|}\sum_{j\in I^L_2}\big(Y_{2j} - \lambda_2 F(X_{2j})\big).\end{aligned}$$

Both estimators are unbiased. The variance-minimizing $(\lambda^\star_1, \lambda^\star_2)$ have closed forms analogous to §4.1 (Appendix C.1). Similarly, we have

$$\mathrm{Var}\left[\widehat{\delta^{PP,\mathrm{Un}}_{1,2}}\right] \le \mathrm{Var}\left[\widehat{\delta^{H,\mathrm{Un}}_{1,2}}\right].$$

5 PPSR: A New Meta-Metric

In this section, we derive a meta-metric for the meta-evaluation question: "how much human annotation does a given metric save?" (§5.1), interpret it as an annotation saving ratio (§5.2), and

[3] Note that for auto-only evaluation, we can have similar confidence intervals and hypothesis tests (Alg. 2 in the appendix), but neither provides statistical guarantees.

[4] Zrnic (2024) proposed using the bootstrap to construct prediction-powered confidence intervals, but it cannot be directly converted into p-values.

contrast it conceptually with existing system-level meta-metrics (§5.3).

5.1 Derivation and Definition

When $U \gg L$, the variance reduction in (5) simplifies cleanly (Appendix D):

$$\frac{\mathrm{Var}[\widehat{\delta^H_{1,2}}] - \mathrm{Var}[\widehat{\delta^{PP}_{1,2}}]}{\mathrm{Var}[\widehat{\delta^H_{1,2}}]} \approx \mathrm{Corr}[Y_1 - Y_2,\ F_k(X_1) - F_k(X_2)]^2 .$$

That is, the relative variance reduction from prediction-powered evaluation over human-only paired evaluation, on a single system pair, is the squared Pearson correlation between human and metric *score differences* on that pair. Averaging over the $\binom{N}{2}$ pairs and plugging in the sample estimator gives the **Prediction-Powered Saving Ratio (PPSR)**:

$$\binom{N}{2}^{-1} \sum_{p=1}^{N-1} \sum_{q=p+1}^{N} r\Big(\{(Y_{pj} - Y_{qj},\ F_k(X_{pj}) - F_k(X_{qj}))\}_{j=1}^{L}\Big)^2$$

PPSR takes values in $[0, 1]$, requires only the labeled paired sample to compute, and is derived from the prediction-powered framework.

5.2 Interpretation

PPSR is, the fraction of human annotation that metric F_k saves relative to human-only paired evaluation while preserving the same confidence-interval width or test power when $U \gg L$ (Appendix D). A PPSR of 0.4 for a metric F_k on a given dataset means that replacing human-only evaluation with PPI using F_k yields the same statistical conclusion with $\approx 40\%$ fewer human-annotated segments on average. PPSR is therefore not just an agreement score: it is a direct prediction of the annotation budget a practitioner will save by adopting F_k in a prediction-powered MT evaluation pipeline.

5.3 Relation to Existing Meta-Metrics

Interestingly, though PPSR is a system-level meta-metric, that is, it reflects the ability of automatic metrics to perform system comparison, it is closely related to segment-level meta-metrics. Compared with existing schemes (Deutsch et al., 2023) including *No Grouping*, *Group by Item*, and *Group by System*, PPSR can be viewed as a segment-level meta-metric under a *Group by System Pair* scheme.

Existing system-level meta-metrics typically aggregate scores at the system level first, by averaging human scores and metric scores over inputs for each system, and then compute a correlation between the resulting N pairs of system-level scores:

$$c\left(\left\{\left(\frac{1}{L}\sum_{j=1}^{L} Y_{ij},\ \frac{1}{L}\sum_{j=1}^{L} F_k(X_{ij})\right)\right\}_{i=1}^{N}\right),$$

where c can be Pearson's r, Spearman's ρ, or Kendall's τ[5]. Because the number of systems N is typically small, these meta-metrics tend to have weak discriminative power and produce unstable rankings.

To address this, Thompson et al. (2024) proposed SPA, which leverages the discrepancy between p-values from human-only and auto-only paired permutation tests as additional signal:

$$\binom{N}{2}^{-1} \sum_{p=1}^{N-1} \sum_{q=p+1}^{N} 1 - \left|p^H_{p,q} - p^A_{p,q}\right|,$$

It has been adopted in the lastest WMT metric shared task (Freitag et al., 2024; Lavie et al., 2025). However, since p-values are not effect sizes or calibrated probabilities, this is best understood as a heuristic measure of agreement.

6 Experiments

We evaluate the framework on six WMT22–24 datasets and organize the experiments around the design choices and meta-evaluation question that structure the paper. After describing the datasets (§6.1), we first verify that prediction-powered evaluation delivers on its basic promise for MT system comparison (§6.2). We then examine the two design axes introduced in §4, parametric vs. non-parametric inference (§6.3) and paired vs. unpaired sampling (§6.4), to quantify the cost of each design choice on real data. Finally, we assess PPSR as a meta-metric and use it to evaluate automatic metrics on the same six datasets, showing that PPSR has higher discriminative power (§6.5) and yields more stable rankings (§6.6) than existing system-level meta-metrics, and we report the metric rankings produced by PPSR (§6.7).

6.1 Datasets

We used the MT Metrics Eval V2 toolkit[6] to obtain all WMT data from 2022 to 2025. After carefully inspecting the data, we selected en-de, en-ru,

[5] Thompson et al. (2024) showed that Pairwise accuracy (PA) and system-level Kendall's τ_a are equivalent up to a linear scaling and shift. However, to handle possible ties, we consistently use τ_b for Kendall's τ.

[6] https://github.com/google-research/mt-metrics-eval

Dataset	Human	#Inputs	#Sys.	#Met.
WMT22 en-de	MQM	1315	14	31
WMT22 en-ru	MQM	1315	15	31
WMT22 zh-en	MQM	1875	14	33
WMT23 en-zh	DA-SQM	1098	15	34
WMT23 ja-en	DA-SQM	1120	17	35
WMT24 cs-uk	ESA	1955	11	25

Table 1: Datasets used in the experiments. Sys. and Met. denote the number of systems and automatic metrics.

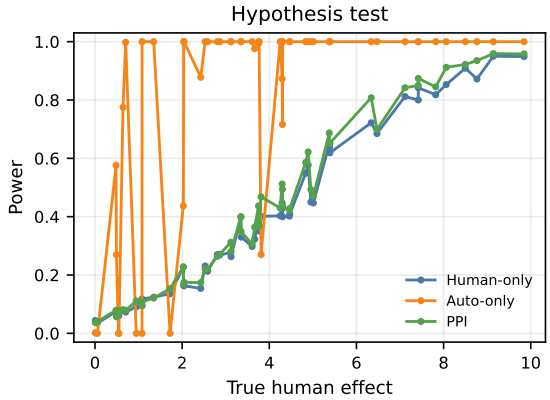

Figure 1: Empirical power curves at $L = 80, U = 800$ on WMT24 cs-uk. The pathological behavior of the auto-only test suggests that even a strong metric such as MetricX exhibits substantial bias for some system pairs. Similar trends are observed for other values of L and other datasets. The full results are shown in Figure 5 and 6 in the appendix.

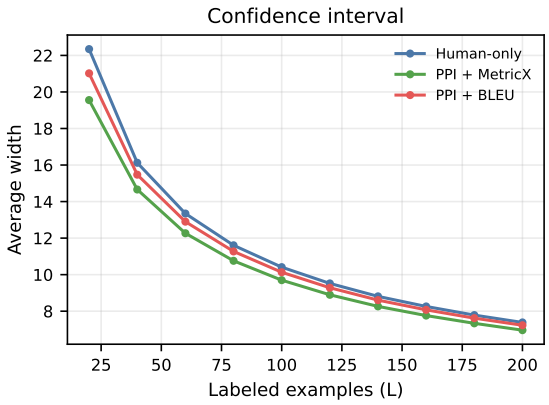

Figure 2: Average two-sided 95% CI width with $U = 800$ on WMT24 cs-uk. A shorter CI means better data efficiency. Results are averaged across system pairs. Auto-only evaluation is excluded because it is biased. The trend is the same across all datasets. Figure 7 in the appendix shows full results.

and zh-en from WMT22; en-zh and ja-en from WMT23; and cs-uk from WMT24. These datasets contain relatively large numbers of source inputs across all systems, making them more suitable for treating the data as a population when evaluating confidence intervals and hypothesis tests. Table 1 summarizes the basic statistics of these datasets. For each input, the human score of each system output and the scores assigned by all automatic metrics are available. We use all available automatic metrics, so our dataset configuration is not identical to that used in the metric shared tasks. For reference-based metrics, we follow the dataset guidelines and use the scores computed with ref-A. For WMT24 cs-uk, we additionally exclude sentinel metrics and three pairs of automatic metrics that produce identical outputs.

6.2 PPI Works for MT System Comparison

Setup. For each system pair, we define the finite population as the full set of paired examples for which human scores are available, and take the mean human score difference over this full set as the population effect size. We then evaluate confidence intervals and hypothesis tests by resampling from this finite population. In each trial, we sample $U + L$ examples without replacement from the full set and randomly split them into L labeled examples, whose human scores are used, and U unlabeled examples, whose human scores are held out. The sampled data are used to construct confidence intervals and conduct hypothesis tests, which are then evaluated against the finite-population effect size defined above. We sweep L from 20 to 200 while fixing $U = 800$. For each (L, U) configuration, we repeat this procedure for 1000 trials. For automatic metrics, we use MetricX: `MetricX-XXL-20` for WMT22, `MetricX-23` for WMT23, and `MetricX-24` for WMT24.

Hypothesis Test. We record the empirical power of each hypothesis test for each system pair, i.e., the percentage of trials in which the one-sided null hypothesis is rejected at $\alpha = 0.05$. Since each system pair has a different true human effect, this produces a power curve. Figure 1 shows the power curve of auto-only tests is markedly different from those of prediction-powered and human-only tests. Its power is abnormally low for some system pairs with a large true human effect due to bias. By contrast, the power curves of human-only and prediction-powered tests align closely, with prediction-powered tests consistently achieving higher power.

Confidence Interval. To better compare data efficiency between human-only and prediction-powered evaluation, we also frame the compari-

son as an interval estimation experiment. We additionally include BLEU as an alternative automatic metric to briefly examine the effect of automatic metrics on efficiency gains. Figure 2 shows that the average width of the prediction-powered CI with MetricX is consistently reduced relative to the human-only baseline, while BLEU yields smaller but still positive reductions. These results also inspired us to design PPSR, selecting the metric that could achieve the most efficiency gains. The empirical coverage rates remain close to the nominal 95% level (see Figure 8 in the appendix).

Summary. Overall, PPI works for MT system comparisons. On the one hand, prediction-powered evaluation is unbiased, which is a clear advantage over auto-only evaluation. On the other hand, it is more data-efficient than human-only evaluation.

6.3 Non-Parametric vs. Parametric Tests

Setup. We follow the experimental setup for hypothesis tests from Section 6.2. For each dataset and each labeled sample size L from 20 to 200, we compute the empirical power for each system pair. We use $B = 1000$ random sign-flip permutations for the paired permutation tests.

Figure 9, 10, 11 in the appendix shows the power curves of the prediction-powered paired permutation test and the human-only paired permutation test. Consistent with the earlier paired Z-test results, we find that their trends are nearly identical, and the prediction-powered test achieves higher power throughout, which is expected.

Figure 3 shows the average power difference across all system pairs from using the paired permutation test instead of the paired Z-test, defined as $\mathrm{Power(Z)} - \mathrm{Power(Perm)}$, for both the human-only and prediction-powered settings. In both settings, the difference is positive. It is largest at very small labeled sample sizes and shrinks toward zero as L increases. We interpret this as evidence that **the parametric approximation is more aggressive, while the paired permutation test provides a useful non-parametric alternative** when the normal approximation is questionable.

6.4 Paired vs. Unpaired Design

When systems are evaluated on disjoint inputs, we must use the unpaired design in Section 4.3. When systems are evaluated on the same source inputs, both paired and unpaired designs are feasible. Which one should we choose? This section answers this question empirically.

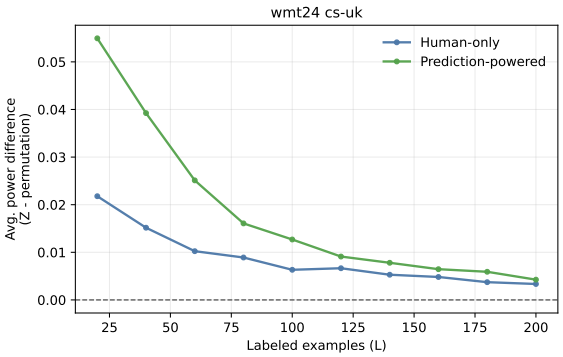

Figure 3: Average power difference on WMT24 cs-uk from using the paired Z-test instead of the paired permutation test. Positive values indicate higher empirical power for the Z-test. The trend is the same across all datasets. Full results for all datasets are shown in Figure 14 in the appendix.

	Human-only		PP	
Dataset	Var. Inc.	Pos.	Var. Inc.	Pos.
WMT22 en-de	1.28	100%	1.11	100%
WMT22 en-ru	1.81	100%	1.69	100%
WMT22 zh-en	1.28	100%	1.03	100%
WMT23 en-zh	0.16	98%	0.11	92%
WMT23 ja-en	0.18	100%	0.12	97%
WMT24 cs-uk	0.56	100%	0.33	100%

Table 2: Paired-versus-unpaired variance comparison for human-only and prediction-powered evaluation. "Var. Inc." is the average relative variance increase from using the unpaired design; "Pos." is the fraction of cases where it is positive (i.e., where pairing reduces variance). For human-only, each case is one system pair; for prediction-powered, each case is one system-pair/metric combination.

Since both $\widehat{\delta_{1,2}^{H,\mathrm{Un}}}$ and $\widehat{\delta_{1,2}^{PP,\mathrm{Un}}}$ are unbiased, we compare them with their paired counterparts in terms of variance. For human-only evaluation, we estimate the relative variance increase $(\widehat{\mathrm{Var}}\left[\widehat{\delta_{1,2}^{H,\mathrm{Un}}}\right] - \widehat{\mathrm{Var}}\left[\widehat{\delta_{1,2}^{H}}\right])/\widehat{\mathrm{Var}}\left[\widehat{\delta_{1,2}^{H}}\right]$ for each system pair in the selected datasets. Similarly, for prediction-powered evaluation, we estimate $(\widehat{\mathrm{Var}}\left[\widehat{\delta_{1,2}^{PP,\mathrm{Un}}}\right] - \widehat{\mathrm{Var}}\left[\widehat{\delta_{1,2}^{PP}}\right])/\widehat{\mathrm{Var}}\left[\widehat{\delta_{1,2}^{PP}}\right]$ for each system pair and metric.

A positive value of the relative variance increase means that the unpaired design has larger estimated variance than the paired design. Table 2 show that this value is positive for nearly all cases, which indicates that, a paired design is usually more data-efficient than an unpaire design. This supports the practical recommendation to **use the paired design for both human-only evaluation and prediction-powered evaluation whenever it is feasible**. How-

Dataset	Distinct values (↑)						Significant comparisons (↑)					
	Max	r	ρ	τ	SPA	PPSR	Max	r	ρ	τ	SPA	PPSR
WMT22 en-de	31	**31**	23	14	**31**	**31**	465	354	290	197	306	**411**
WMT22 en-ru	31	**31**	24	17	**31**	**31**	465	386	311	290	340	**404**
WMT22 zh-en	31	**31**	27	18	**31**	**31**	465	408	350	318	358	**422**
WMT23 en-zh	34	**34**	26	20	**34**	**34**	561	490	423	405	440	**509**
WMT23 ja-en	35	**35**	22	13	**35**	**35**	595	395	271	272	338	**501**
WMT24 cs-uk	25	**25**	17	14	**25**	**25**	300	233	208	197	225	**247**

Table 3: Discriminative power of system-level meta-metrics. Max gives the largest possible value for each criterion. Bold indicates the highest observed value in each row.

ever, this effect in prediction-powered evaluation is weaker than in the human-only cases. Please see Appendix C.2 for a more in-depth analysis of this phenomenon.

6.5 PPSR Has Higher Discriminative Power

A meta-metric with higher discriminative power is more likely to distinguish between automatic metrics. This is desirable in meta-evaluation, as an excess of ties makes it comparison difficult.

We evaluate discriminative power in two ways. First, we count how many distinct values each meta-metric assigns to the automatic metrics in a dataset. Second, for every pair of automatic metrics, we follow the practice of Thompson et al. (2024) to apply the PERM-INPUTS permutation test (Deutsch et al., 2021) with $B = 1000$ and count how many pairwise differences are statistically significant at $p \leq 0.05$. Table 3 summarizes the results. We find that PPSR achieves the highest discriminative power among all system-level meta-metrics on each dataset. By comparison, SPA has higher discriminative power than system-level ρ and τ, but lower than system-level r.

6.6 PPSR Produces More Stable Rankings

Another dimension to assess meta-metrics is whether the induced ranking of automatic metrics remains stable when only a subset of data is available. This matters because meta-evaluation studies often compare many metrics on a limited test set.

We use resampling to assess ranking stability. For each dataset and each meta-metric, we first obtain the ranking of automatic metrics using the full dataset. We then sample L inputs without replacement (keeping the system set fixed), recompute each meta-metric's ranking on the sampled data, and measure Kendall's τ between the sampled ranking and the full-data ranking. This procedure is repeated 1000 times for each $L \in \{100, 200, \ldots, 1000\}$, and we report the average

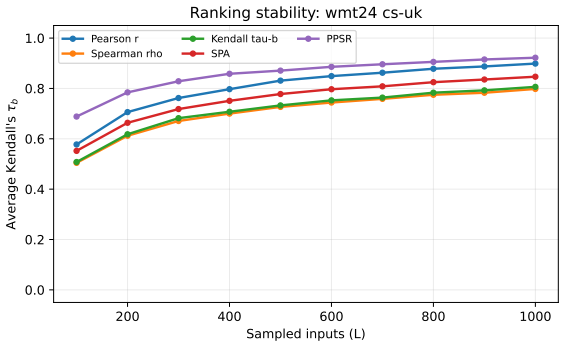

Figure 4: Ranking stability of system-level meta-metrics on WMT24 cs-uk. Higher values indicate more stable metric rankings. Note that for meta-evaluation, only labeled examples are used and the unlabeled examples are unused. The trend is the same across all datasets. Full results for all datasets are shown in Figure 15 in the appendix.

Kendall's τ. Figure 4 shows PPSR achieves the highest average ranking stability across input size.

6.7 Metric Rankings Produced By PPSR

Tables 5–10 in the appendix report the score and rank of every automatic metric under PPSR and other system-level meta-metrics. The top-ranked automatic metric is not generally consistent across meta-metrics: PPSR agrees with all four comparison meta-metrics on the top metric (`mqm.unb`) only for WMT22 en-ru. This supports the interpretation that PPSR emphasizes a related but distinct property of automatic metrics.

7 Conclusion

We demonstrate the advantages of prediction-powered evaluation over both human-only and auto-only evaluation using WMT data. We further provide recommendations for two design axes: parametric versus non-parametric methods, and paired versus unpaired designs. Finally, building on prediction-powered evaluation, we derive PPSR, a new meta-metric with greater discriminative power and improved ranking stability.

Limitations

For simplicity, we do not model inter-annotator disagreement as an additional source of variance in human scores, as considered in prior work (Chaganty et al., 2018; Wei and Jia, 2021).

We also assume that system-level scores are obtained by averaging over inputs. As a result, our methods may not apply directly to the automatic metrics that compute system-level scores using different aggregation procedures.

Finally, our experiments focus on machine translation. The applicability of our methods to other domains and tasks remains to be further validated.

Ethical Considerations

The MT Metrics Eval V2 toolkit is licensed under Apache-2.0, and our use of it complies with the terms of the license. AI-based writing tools were used to assist with language polishing.

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondrej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussà, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, and 17 others. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation, WMT@EMNLP 2021, Online Event, November 10-11, 2021*, pages 1–88. Association for Computational Linguistics.

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. 2023a. Prediction-powered inference. *Science*, 382(6671):669–674.

Anastasios N. Angelopoulos, John C. Duchi, and Tijana Zrnic. 2023b. PPI++: efficient prediction-powered inference. *CoRR*, abs/2311.01453.

Loïc Barrault, Magdalena Biesialska, Ondrej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubesic, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, and 2 others. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation, WMT@EMNLP 2020, Online, November 19-20, 2020*, pages 1–55. Association for Computational Linguistics.

Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno-Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana L. Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, and 2 others. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation, WMT 2016, colocated with ACL 2016, August 11-12, Berlin, Germany*, pages 131–198. The Association for Computer Linguistics.

Pierre Boyeau, Anastasios Nikolas Angelopoulos, Tianle Li, Nir Yosef, Jitendra Malik, and Michael I. Jordan. 2025. Autoeval done right: Using synthetic data for model evaluation. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, Proceedings of Machine Learning Research. PMLR / OpenReview.net.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation, WMT@NAACL-HLT 2012, June 7-8, 2012, Montréal, Canada*, pages 10–51. The Association for Computer Linguistics.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 9263–9274. Association for Computational Linguistics.

Arun Chaganty, Stephen Mussmann, and Percy Liang. 2018. The price of debiasing automatic metrics in natural language evalaution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653, Melbourne, Australia. Association for Computational Linguistics.

Ivi Chatzi, Eleni Straitouri, Suhas Thejaswi, and Manuel Gomez Rodriguez. 2024. Prediction-powered ranking of large language models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Yiqun T. Chen, Moran Guo, and Shengy Li. 2026a. Power analysis for prediction-powered inference. *CoRR*, abs/2603.16041.

Yiqun T. Chen, Sizhu Lu, Sijia Li, Moran Guo, and Shengyi Li. 2026b. Efficient inference for noisy llm-as-a-judge evaluation. *CoRR*, abs/2601.05420.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2021. A statistical analysis of summarization evaluation metrics using resampling methods. *Trans. Assoc. Comput. Linguistics*, 9:1132–1146.

Daniel Deutsch, George F. Foster, and Markus Freitag. 2023. Ties matter: Meta-evaluating modern metrics with pairwise accuracy and tie calibration. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 12914–12929. Association for Computational Linguistics.

Colten DiIanni and Daniel Deutsch. 2025. Don't sweat the small stuff: Segment-level meta-evaluation based on pairwise difference correlation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 25062–25070. Association for Computational Linguistics.

Florian E. Dorner, Vivian Yvonne Nastl, and Moritz Hardt. 2025. Limits to scalable evaluation at the frontier: LLM as judge won't beat twice the data. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1383–1392. Association for Computational Linguistics.

Chen Feng, Minghe Shen, Ananth Balashankar, Carsten Gerner-Beuerle, and Miguel R. D. Rodrigues. 2026. Noisy but valid: Robust statistical evaluation of llms with imperfect judges. *CoRR*, abs/2601.20913.

Adam Fisch, Joshua Maynez, R. Alex Hofer, Bhuwan Dhingra, Amir Globerson, and William W. Cohen. 2024. Stratified prediction-powered inference for effective hybrid evaluation of language models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frédéric 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, WMT 2024, Miami, FL, USA, November 15-16, 2024*, pages 47–81. Association for Computational Linguistics.

Mingqi Gao, Xinyu Hu, Li Lin, and Xiaojun Wan. 2025. Analyzing and evaluating correlation measures in NLG meta-evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 2199–2222. Association for Computational Linguistics.

Yicheng Gao, Gonghan Xu, Zhe Wang, and Arman Cohan. 2024. Bayesian calibration of win rate estimation with LLM evaluators. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 4757–4769. Association for Computational Linguistics.

Yvette Graham, Nitika Mathur, and Timothy Baldwin. 2014. Randomized significance tests in machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation, WMT@ACL 2014, June 26-27, 2014, Baltimore, Maryland, USA*, pages 266–274. The Association for Computer Linguistics.

Luke Guerdan, Justin Whitehouse, Kimberly Le Truong, Kenneth Holstein, and Zhiwei Steven Wu. 2025. Doubly-robust llm-as-a-judge: Externally valid estimation with imperfect personas. *CoRR*, abs/2509.22957.

David M. Howcroft and Verena Rieser. 2021. What happens if you treat ordinal ratings as interval data? human evaluations in NLP are even more underpowered than you think. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 8932–8939. Association for Computational Linguistics.

Wenlong Ji, Lihua Lei, and Tijana Zrnic. 2025. Predictions as surrogates: Revisiting surrogate outcomes in the age of AI. *CoRR*, abs/2501.09731.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Christof Monz, Kenton Murray, and 10 others. 2025. Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413, Suzhou, China. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondrej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popovic, and 3 others. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation, WMT 2024, Miami, FL, USA, November 15-16, 2024*, pages 1–46. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondrej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme

Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, and 2 others. 2023. Findings of the 2023 conference on machine translation (WMT23): llms are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation, WMT 2023, Singapore, December 6-7, 2023*, pages 1–42. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondrej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popovic. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation, WMT 2022, Abu Dhabi, United Arab Emirates (Hybrid), December 7-8, 2022*, pages 1–45. Association for Computational Linguistics.

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, WMT@EMNLP 2021, Online Event, November 10-11, 2021*, pages 478–494. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing , EMNLP 2004, A meeting of SIGDAT, a Special Interest Group of the ACL, held in conjunction with ACL 2004, 25-26 July 2004, Barcelona, Spain*, pages 388–395. ACL.

Alon Lavie, Greg Hanneman, Sweta Agrawal, Diptesh Kanojia, Chi-Kiu Lo, Vilém Zouhar, Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, Archchana Sindhujan, 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*, pages 436–483, Suzhou, China. Association for Computational Linguistics.

Chungpa Lee, Thomas Zeng, Jongwon Jeong, Jy-yong Sohn, and Kangwook Lee. 2025. How to correctly report llm-as-a-judge evaluations. *CoRR*, abs/2511.21140.

Qingsong Ma, Johnny Wei, Ondrej 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, WMT 2019, Florence, Italy, August 1-2, 2019 - Volume 2: Shared Task Papers, Day 1*, pages 62–90. Association for Computational Linguistics.

Matous Machácek and Ondrej Bojar. 2013. Results of the WMT13 metrics shared task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation, WMT@ACL 2013, August 8-9, 2013, Sofia, Bulgaria*. The Association for Computer Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020a. 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, ACL 2020, Online, July 5-10, 2020*, pages 4984–4997. Association for Computational Linguistics.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondrej Bojar. 2020b. Results of the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation, WMT@EMNLP 2020, Online, November 19-20, 2020*, pages 688–725. Association for Computational Linguistics.

Stefano Perrella, Lorenzo Proietti, Pere-Lluís Huguet Cabot, Edoardo Barba, and Roberto Navigli. 2024. Beyond correlation: Interpretable evaluation of machine translation metrics. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 20689–20714. Association for Computational Linguistics.

Walter J. Rogan and Beth C. Gladen. 1978. Estimating prevalence from the results of a screening test. *American journal of epidemiology*, 107 1:71–6.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. ARES: an automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 338–354. Association for Computational Linguistics.

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, WMT 2024, Miami, FL, USA, November 15-16, 2024*, pages 1222–1234. Association for Computational Linguistics.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Comput. Speech Lang.*, 67:101151.

Johnny Tian-Zheng Wei and Robin Jia. 2021. The statistical advantage of automatic NLG metrics at the system level. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference*

on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021, pages 6840–6854. Association for Computational Linguistics.

Zhaoyi Zhou, Yuda Song, and Andrea Zanette. 2025. Accelerating unbiased LLM evaluation via synthetic feedback. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, Proceedings of Machine Learning Research. PMLR / OpenReview.net.

Tijana Zrnic. 2024. A note on the prediction-powered bootstrap. *CoRR*, abs/2405.18379.

A $\mathbb{E}[\widehat{\delta_{1,2}^{PP}}]$ and $\mathrm{Var}[\widehat{\delta_{1,2}^{PP}}]$

Denote $\mu_i = \mathbb{E}[Y_i]$ and $\nu_i = \mathbb{E}[F(X_i)]$.

$$
\begin{aligned}
\mathbb{E}[\widehat{\delta_{1,2}^{PP}}] &= \mathbb{E}\left[\frac{1}{U}\sum_{j=L+1}^{L+U}\lambda\big(F(X_{1j})-F(X_{2j})\big)+\frac{1}{L}\sum_{j=1}^{L}\big[(Y_{1j}-Y_{2j})-\lambda\big(F(X_{1j})-F(X_{2j})\big)\big]\right]\\
&= \frac{\lambda}{U}\sum_{j=L+1}^{L+U}\mathbb{E}\left[F(X_{1j})-F(X_{2j})\right]+\frac{1}{L}\sum_{j=1}^{L}\mathbb{E}\left[Y_{1j}-Y_{2j}\right]-\frac{\lambda}{L}\sum_{j=1}^{L}\mathbb{E}\left[F(X_{1j})-F(X_{2j})\right]\\
&= \frac{\lambda}{U}\sum_{j=L+1}^{L+U}\mathbb{E}\left[F(X_{1j})\right]-\frac{\lambda}{U}\sum_{j=L+1}^{L+U}\mathbb{E}\left[F(X_{2j})\right]+\frac{1}{L}\sum_{j=1}^{L}\mathbb{E}\left[Y_{1j}\right]\\
&\quad -\frac{1}{L}\sum_{j=1}^{L}\mathbb{E}\left[Y_{2j}\right]-\frac{\lambda}{L}\sum_{j=1}^{L}\mathbb{E}\left[F(X_{1j})\right]+\frac{\lambda}{L}\sum_{j=1}^{L}\mathbb{E}\left[F(X_{2j})\right]\\
&= \lambda\nu_1-\lambda\nu_2+\mu_1-\mu_2-\lambda\nu_1+\lambda\nu_2\\
&= \mu_1-\mu_2\\
&= \delta_{1,2}.
\end{aligned}
$$

Therefore, $\widehat{\delta_{1,2}^{PP}}$ is an unbiased estimator of $\delta_{1,2}$.

$$
\begin{aligned}
\mathrm{Var}[\widehat{\delta_{1,2}^{PP}}] &= \mathrm{Var}\left[\frac{1}{U}\sum_{j=L+1}^{L+U}\lambda\big(F(X_{1j})-F(X_{2j})\big)+\frac{1}{L}\sum_{j=1}^{L}\big[(Y_{1j}-Y_{2j})-\lambda\big(F(X_{1j})-F(X_{2j})\big)\big]\right]\\
&= \mathrm{Var}\left[\frac{1}{U}\sum_{j=L+1}^{L+U}\lambda\big(F(X_{1j})-F(X_{2j})\big)\right]+\mathrm{Var}\left[\frac{1}{L}\sum_{j=1}^{L}\big[(Y_{1j}-Y_{2j})-\lambda\big(F(X_{1j})-F(X_{2j})\big)\big]\right]\\
&= \frac{\lambda^2}{U}\mathrm{Var}\big[F(X_{1j})-F(X_{2j})\big]+\frac{1}{L}\mathrm{Var}\left[(Y_{1j}-Y_{2j})-\lambda\big(F(X_{1j})-F(X_{2j})\big)\right]\\
&= \frac{\lambda^2}{U}\mathrm{Var}\big[F(X_{1j})-F(X_{2j})\big]+\frac{1}{L}\Big(\mathrm{Var}[Y_{1j}-Y_{2j}]+\lambda^2\mathrm{Var}\big[F(X_{1j})-F(X_{2j})\big]\\
&\qquad -2\lambda\mathrm{Cov}\big[Y_{1j}-Y_{2j},\,F(X_{1j})-F(X_{2j})\big]\Big)\\
&= \lambda^2\left(\frac{1}{U}+\frac{1}{L}\right)\mathrm{Var}\big[F(X_{1j})-F(X_{2j})\big]-\frac{2\lambda}{L}\mathrm{Cov}\big[Y_{1j}-Y_{2j},\,F(X_{1j})-F(X_{2j})\big]+\frac{1}{L}\mathrm{Var}[Y_{1j}-Y_{2j}]\\
&= \lambda^2\left(\frac{1}{U}+\frac{1}{L}\right)\mathrm{Var}\big[F(X_1)-F(X_2)\big]-\frac{2\lambda}{L}\mathrm{Cov}\big[Y_1-Y_2,\,F(X_1)-F(X_2)\big]+\frac{1}{L}\mathrm{Var}[Y_1-Y_2].
\end{aligned}
$$

This is an upward-opening quadratic function of λ. When

$$
\lambda^\star = \frac{\mathrm{Cov}\big[Y_1-Y_2,\,F(X_1)-F(X_2)\big]}{\left(1+\frac{L}{U}\right)\mathrm{Var}\big[F(X_1)-F(X_2)\big]},
$$

the variance $\mathrm{Var}[\widehat{\delta_{1,2}^{PP}}]$ attains its minimum value. Substituting $\lambda^\star$ into the variance gives

$$
\begin{aligned}
&\mathrm{Var}[\widehat{\delta_{1,2}^{PP}}]\\
&= \frac{1}{L}\mathrm{Var}[Y_1-Y_2]-\frac{U\left(\mathrm{Cov}\big[Y_1-Y_2,\,F(X_1)-F(X_2)\big]\right)^2}{L(L+U)\mathrm{Var}\big[F(X_1)-F(X_2)\big]}\\
&= \mathrm{Var}[\widehat{\delta_{1,2}^{H}}]-\frac{U\left(\mathrm{Cov}\big[Y_1-Y_2,\,F(X_1)-F(X_2)\big]\right)^2}{L(L+U)\mathrm{Var}\big[F(X_1)-F(X_2)\big]}\\
&\le \mathrm{Var}[\widehat{\delta_{1,2}^{H}}]
\end{aligned}
$$

B Proof of Asymptotic Confidence Interval Validity

We now show that the confidence interval based on $\widehat{\delta_{1,2}^{PP}}$ is asymptotically valid for any fixed λ. The human-only estimator is obtained as the special case $\lambda = 0$.

Let

$$
D_j^Y = Y_{1j} - Y_{2j}, \qquad D_j^F = F(X_{1j}) - F(X_{2j}).
$$

Then the prediction-powered estimator can be written as

$$
\widehat{\delta_{1,2}^{PP}} = \frac{1}{U}\sum_{j=L+1}^{L+U}\lambda D_j^F + \frac{1}{L}\sum_{j=1}^{L}\left(D_j^Y - \lambda D_j^F\right).
$$

From Appendix A, we have

$$
\mathbb{E}\left[\widehat{\delta_{1,2}^{PP}}\right] = \delta_{1,2},
$$

and

$$
\mathrm{Var}\left[\widehat{\delta_{1,2}^{PP}}\right] = \frac{\lambda^2}{U}\mathrm{Var}[D_j^F] + \frac{1}{L}\mathrm{Var}\big[D_j^Y - \lambda D_j^F\big].
$$

Assume that the labeled examples $\{(D_j^Y, D_j^F)\}_{j=1}^{L}$ are i.i.d., the unlabeled examples $\{D_j^F\}_{j=L+1}^{L+U}$ are i.i.d., and the labeled and unlabeled samples are independent. Also assume that the relevant second moments are finite and that $L, U \to \infty$.

By the central limit theorem,

$$
\frac{\frac{1}{U}\sum_{j=L+1}^{L+U}\lambda D_j^F - \lambda\mathbb{E}[D_j^F]}{\sqrt{\frac{\lambda^2}{U}\mathrm{Var}[D_j^F]}} \xrightarrow{d} \mathcal{N}(0,1),
$$

and

$$
\frac{\frac{1}{L}\sum_{j=1}^{L}(D_j^Y - \lambda D_j^F) - \mathbb{E}[D_j^Y - \lambda D_j^F]}{\sqrt{\frac{1}{L}\mathrm{Var}[D_j^Y - \lambda D_j^F]}} \xrightarrow{d} \mathcal{N}(0,1).
$$

Since the two sample averages are independent, their sum is also asymptotically normal. Therefore,

$$
\frac{\widehat{\delta_{1,2}^{PP}} - \delta_{1,2}}{\sqrt{\mathrm{Var}\left[\widehat{\delta_{1,2}^{PP}}\right]}} \xrightarrow{d} \mathcal{N}(0,1).
$$

It follows that

$$
\widehat{\delta_{1,2}^{PP}} \pm z_{\alpha/2}\sqrt{\mathrm{Var}\left[\widehat{\delta_{1,2}^{PP}}\right]}
$$

is an asymptotic $100(1-\alpha)\%$ confidence interval for $\delta_{1,2}$, where $z_{\alpha/2}$ denotes the upper $\alpha/2$ quantile of the standard normal distribution.

The human-only estimator is a special case of the prediction-powered estimator. Specifically, when $\lambda = 0$,

$$\widehat{\delta_{1,2}^{PP}} = \frac{1}{L}\sum_{j=1}^{L} D_j^Y = \widehat{\delta_{1,2}^{H}}.$$

Thus,

$$\mathrm{Var}\left[\widehat{\delta_{1,2}^{PP}}\right] = \frac{1}{L}\mathrm{Var}[D_j^Y] = \mathrm{Var}\left[\widehat{\delta_{1,2}^{H}}\right].$$

Hence the same asymptotic confidence interval form applies to both human-only evaluation and prediction-powered evaluation:

$$\widehat{\delta_{1,2}^{g}} \pm z_{\alpha/2}\sqrt{\mathrm{Var}\left[\widehat{\delta_{1,2}^{g}}\right]}, \qquad g \in \{H, PP\}.$$

In practice, the population variance is replaced by a consistent variance estimator. By Slutsky's theorem,

$$\frac{\widehat{\delta_{1,2}^{g}} - \delta_{1,2}}{\sqrt{\widehat{\mathrm{Var}}\left[\widehat{\delta_{1,2}^{g}}\right]}} \xrightarrow{d} \mathcal{N}(0,1), \qquad g \in \{H, PP\}.$$

Therefore, the feasible asymptotic confidence interval is

$$\widehat{\delta_{1,2}^{g}} \pm z_{\alpha/2}\sqrt{\widehat{\mathrm{Var}}\left[\widehat{\delta_{1,2}^{g}}\right]}, \qquad g \in \{H, PP\}.$$

C Paired vs. Unpaired Design

C.1 $\mathbb{E}[\widehat{\delta_{1,2}^{PP,\mathrm{Un}}}]$ and $\mathrm{Var}[\widehat{\delta_{1,2}^{PP,\mathrm{Un}}}]$

Let

$$L_i = |I_i^L|, \qquad U_i = |I_i^U|,$$

and denote

$$\mu_i = \mathbb{E}[Y_i], \qquad \nu_i = \mathbb{E}[F(X_i)].$$

Under the unpaired design, we assume that samples from different systems are independent, and that the labeled and unlabeled samples for each system are drawn from the same system-specific distribution.

$$\begin{aligned}
\mathbb{E}\left[\widehat{\delta_{1,2}^{PP,\mathrm{Un}}}\right] &= \mathbb{E}\Bigg[\frac{1}{U_1}\sum_{j\in I_1^U}\lambda_1 F(X_{1j}) + \frac{1}{L_1}\sum_{j\in I_1^L}\left(Y_{1j} - \lambda_1 F(X_{1j})\right) \\
&\qquad - \frac{1}{U_2}\sum_{j\in I_2^U}\lambda_2 F(X_{2j}) - \frac{1}{L_2}\sum_{j\in I_2^L}\left(Y_{2j} - \lambda_2 F(X_{2j})\right)\Bigg] \\
&= \frac{\lambda_1}{U_1}\sum_{j\in I_1^U}\mathbb{E}[F(X_{1j})] + \frac{1}{L_1}\sum_{j\in I_1^L}\mathbb{E}[Y_{1j}] - \frac{\lambda_1}{L_1}\sum_{j\in I_1^L}\mathbb{E}[F(X_{1j})] \\
&\qquad - \frac{\lambda_2}{U_2}\sum_{j\in I_2^U}\mathbb{E}[F(X_{2j})] - \frac{1}{L_2}\sum_{j\in I_2^L}\mathbb{E}[Y_{2j}] + \frac{\lambda_2}{L_2}\sum_{j\in I_2^L}\mathbb{E}[F(X_{2j})] \\
&= \lambda_1\nu_1 + \mu_1 - \lambda_1\nu_1 - \lambda_2\nu_2 - \mu_2 + \lambda_2\nu_2 \\
&= \mu_1 - \mu_2 \\
&= \delta_{1,2}.
\end{aligned}$$

Therefore, $\widehat{\delta_{1,2}^{PP,\mathrm{Un}}}$ is an unbiased estimator of $\delta_{1,2}$.

$$\begin{aligned}
\mathrm{Var}\left[\widehat{\delta_{1,2}^{PP,\mathrm{Un}}}\right] &= \mathrm{Var}\Bigg[\frac{1}{U_1}\sum_{j\in I_1^U}\lambda_1 F(X_{1j}) + \frac{1}{L_1}\sum_{j\in I_1^L}\left(Y_{1j} - \lambda_1 F(X_{1j})\right) \\
&\qquad - \frac{1}{U_2}\sum_{j\in I_2^U}\lambda_2 F(X_{2j}) - \frac{1}{L_2}\sum_{j\in I_2^L}\left(Y_{2j} - \lambda_2 F(X_{2j})\right)\Bigg] \\
&= \mathrm{Var}\left[\frac{1}{U_1}\sum_{j\in I_1^U}\lambda_1 F(X_{1j})\right] + \mathrm{Var}\left[\frac{1}{L_1}\sum_{j\in I_1^L}\left(Y_{1j} - \lambda_1 F(X_{1j})\right)\right] \\
&\qquad + \mathrm{Var}\left[\frac{1}{U_2}\sum_{j\in I_2^U}\lambda_2 F(X_{2j})\right] + \mathrm{Var}\left[\frac{1}{L_2}\sum_{j\in I_2^L}\left(Y_{2j} - \lambda_2 F(X_{2j})\right)\right] \\
&= \frac{\lambda_1^2}{U_1}\mathrm{Var}[F(X_1)] + \frac{1}{L_1}\mathrm{Var}\left[Y_1 - \lambda_1 F(X_1)\right] \\
&\qquad + \frac{\lambda_2^2}{U_2}\mathrm{Var}[F(X_2)] + \frac{1}{L_2}\mathrm{Var}\left[Y_2 - \lambda_2 F(X_2)\right] \\
&= \frac{\lambda_1^2}{U_1}\mathrm{Var}[F(X_1)] + \frac{1}{L_1}\left(\mathrm{Var}[Y_1] + \lambda_1^2\mathrm{Var}[F(X_1)] - 2\lambda_1\mathrm{Cov}[Y_1, F(X_1)]\right) \\
&\qquad + \frac{\lambda_2^2}{U_2}\mathrm{Var}[F(X_2)] + \frac{1}{L_2}\left(\mathrm{Var}[Y_2] + \lambda_2^2\mathrm{Var}[F(X_2)] - 2\lambda_2\mathrm{Cov}[Y_2, F(X_2)]\right) \\
&= \lambda_1^2\left(\frac{1}{U_1} + \frac{1}{L_1}\right)\mathrm{Var}[F(X_1)] - \frac{2\lambda_1}{L_1}\mathrm{Cov}[Y_1, F(X_1)] + \frac{1}{L_1}\mathrm{Var}[Y_1] \\
&\qquad + \lambda_2^2\left(\frac{1}{U_2} + \frac{1}{L_2}\right)\mathrm{Var}[F(X_2)] - \frac{2\lambda_2}{L_2}\mathrm{Cov}[Y_2, F(X_2)] + \frac{1}{L_2}\mathrm{Var}[Y_2].
\end{aligned}$$

This variance is a sum of two upward-opening quadratic functions, one in λ_1 and the other in λ_2. Therefore, it is minimized when

$$\lambda_i^\star = \frac{\mathrm{Cov}[Y_i, F(X_i)]}{\left(1 + \frac{L_i}{U_i}\right)\mathrm{Var}[F(X_i)]}, \qquad i \in \{1, 2\}.$$

Substituting $\lambda_1^\star$ and $\lambda_2^\star$ into the variance gives

$$\begin{aligned}
\mathrm{Var}\left[\widehat{\delta_{1,2}^{PP,\mathrm{Un}}}\right] &= \frac{1}{L_1}\mathrm{Var}[Y_1] - \frac{U_1\left(\mathrm{Cov}[Y_1, F(X_1)]\right)^2}{L_1(L_1+U_1)\mathrm{Var}[F(X_1)]} \\
&\qquad + \frac{1}{L_2}\mathrm{Var}[Y_2] - \frac{U_2\left(\mathrm{Cov}[Y_2, F(X_2)]\right)^2}{L_2(L_2+U_2)\mathrm{Var}[F(X_2)]} \\
&= \mathrm{Var}\left[\widehat{\delta_{1,2}^{H,\mathrm{Un}}}\right] - \frac{U_1\left(\mathrm{Cov}[Y_1, F(X_1)]\right)^2}{L_1(L_1+U_1)\mathrm{Var}[F(X_1)]} - \frac{U_2\left(\mathrm{Cov}[Y_2, F(X_2)]\right)^2}{L_2(L_2+U_2)\mathrm{Var}[F(X_2)]} \\
&\le \mathrm{Var}\left[\widehat{\delta_{1,2}^{H,\mathrm{Un}}}\right].
\end{aligned}$$

C.2 Difference between unpaired variance and paired variance

Assume that the paired and unpaired estimators are compared under the same sample sizes, namely

$$L_1 = L_2 = L, \qquad U_1 = U_2 = U.$$

The paired variance of human-only evaluation is

$$\mathrm{Var}\left[\widehat{\delta^{H}_{1,2}}\right] = \frac{1}{L}\mathrm{Var}[Y_1 - Y_2]$$
$$= \frac{1}{L}\left(\mathrm{Var}[Y_1] + \mathrm{Var}[Y_2] - 2\mathrm{Cov}[Y_1, Y_2]\right).$$

By contrast, the unpaired variance is

$$\mathrm{Var}\left[\widehat{\delta^{H,\mathrm{Un}}_{1,2}}\right] = \frac{1}{L}\left(\mathrm{Var}[Y_1] + \mathrm{Var}[Y_2]\right).$$

Thus

$$\mathrm{Var}[\widehat{\delta^{H,\mathrm{Un}}_{1,2}}] - \mathrm{Var}[\widehat{\delta^{H}_{1,2}}] = \frac{2}{L}\mathrm{Cov}[Y_1, Y_2].$$

If $\mathrm{Cov}[Y_1, Y_2] > 0$, i.e., if the human scores of S_1 and S_2 are positively correlated, then using an unpaired design leads to a larger variance.

When it comes to prediction-powered evaluation, the difference between unpaired variance and paired variance is more complex because the automatic metric F is involved. The unpaired variance of prediction-powered evaluation is

$$\mathrm{Var}\left[\widehat{\delta^{PP,\mathrm{Un}}_{1,2}}\right] = \frac{1}{L}\mathrm{Var}[Y_1] - \frac{U\left(\mathrm{Cov}[Y_1, F(X_1)]\right)^2}{L(L+U)\mathrm{Var}[F(X_1)]} + \frac{1}{L}\mathrm{Var}[Y_2] - \frac{U\left(\mathrm{Cov}[Y_2, F(X_2)]\right)^2}{L(L+U)\mathrm{Var}[F(X_2)]}.$$

The paired variance is

$$\mathrm{Var}\left[\widehat{\delta^{PP}_{1,2}}\right] = \frac{1}{L}\mathrm{Var}[Y_1 - Y_2] - \frac{U\left(\mathrm{Cov}\left[Y_1 - Y_2,\ F(X_1) - F(X_2)\right]\right)^2}{L(L+U)\mathrm{Var}\left[F(X_1) - F(X_2)\right]}.$$

Therefore,

$$\mathrm{Var}\left[\widehat{\delta^{PP,\mathrm{Un}}_{1,2}}\right] - \mathrm{Var}\left[\widehat{\delta^{PP}_{1,2}}\right]$$
$$= \frac{1}{L}\mathrm{Var}[Y_1] + \frac{1}{L}\mathrm{Var}[Y_2] - \frac{1}{L}\mathrm{Var}[Y_1 - Y_2] - \frac{U\left(\mathrm{Cov}[Y_1, F(X_1)]\right)^2}{L(L+U)\mathrm{Var}[F(X_1)]} - \frac{U\left(\mathrm{Cov}[Y_2, F(X_2)]\right)^2}{L(L+U)\mathrm{Var}[F(X_2)]} + \frac{U\left(\mathrm{Cov}\left[Y_1 - Y_2,\ F(X_1) - F(X_2)\right]\right)^2}{L(L+U)\mathrm{Var}\left[F(X_1) - F(X_2)\right]}.$$

As in $\mathrm{Var}[\widehat{\delta^{H,\mathrm{Un}}_{1,2}}] - \mathrm{Var}[\widehat{\delta^{H}_{1,2}}]$, we have

$$\mathrm{Var}[Y_1] + \mathrm{Var}[Y_2] - \mathrm{Var}[Y_1 - Y_2] = 2\mathrm{Cov}[Y_1, Y_2].$$

By the definition of correlation, we have

$$\frac{\left(\mathrm{Cov}[A, B]\right)^2}{\mathrm{Var}[B]} = \mathrm{Corr}[A, B]^2\mathrm{Var}[A].$$

Dataset	$\lvert\widehat{\mathrm{Corr}_{\delta}}\rvert$	$\lvert\widehat{\mathrm{Corr}_{\mathrm{sep}}}\rvert$	Smaller
WMT22 en-de	0.227	0.327	92.6%
WMT22 en-ru	0.218	0.282	79.4%
WMT22 zh-en	0.169	0.333	93.6%
WMT23 en-zh	0.244	0.288	69.7%
WMT23 ja-en	0.215	0.291	89.7%
WMT24 cs-uk	0.246	0.427	96.1%

Table 4: Correlation diagnostics for the prediction-powered paired-versus-unpaired comparison. Here $\widehat{\mathrm{Corr}_{\delta}} = \widehat{\mathrm{Corr}}[Y_1 - Y_2, F(X_1) - F(X_2)]$, and $\widehat{\mathrm{Corr}_{\mathrm{sep}}} = \frac{1}{2}(\widehat{\mathrm{Corr}}[Y_1, F(X_1)] + \widehat{\mathrm{Corr}}[Y_2, F(X_2)])$. Mean values across system pairs and automatic metrics are shown. "Smaller" is the percentage of cases where $\lvert\widehat{\mathrm{Corr}_{\delta}}\rvert < \lvert\widehat{\mathrm{Corr}_{\mathrm{sep}}}\rvert$.

Applying this to each covariance term gives

$$\mathrm{Var}\left[\widehat{\delta^{PP,\mathrm{Un}}_{1,2}}\right] - \mathrm{Var}\left[\widehat{\delta^{PP}_{1,2}}\right]$$
$$= \frac{2}{L}\mathrm{Cov}[Y_1, Y_2] + \frac{U}{L(L+U)}\Big(\mathrm{Corr}\left[Y_1 - Y_2,\ F(X_1) - F(X_2)\right]^2\mathrm{Var}[Y_1 - Y_2] - \mathrm{Corr}[Y_1, F(X_1)]^2\mathrm{Var}[Y_1] - \mathrm{Corr}[Y_2, F(X_2)]^2\mathrm{Var}[Y_2]\Big)$$
$$= \frac{1}{L}\Bigg(2\mathrm{Cov}[Y_1, Y_2] + \frac{U}{L+U}\Big(\mathrm{Corr}\left[Y_1 - Y_2,\ F(X_1) - F(X_2)\right]^2\mathrm{Var}[Y_1 - Y_2] - \mathrm{Corr}[Y_1, F(X_1)]^2\mathrm{Var}[Y_1] - \mathrm{Corr}[Y_2, F(X_2)]^2\mathrm{Var}[Y_2]\Big)\Bigg).$$

If we further uses the approximation $\frac{U}{L+U} \approx 1$, for example when $U \gg L$, then this simplifies to

$$\mathrm{Var}\left[\widehat{\delta^{PP,\mathrm{Un}}_{1,2}}\right] - \mathrm{Var}\left[\widehat{\delta^{PP}_{1,2}}\right]$$
$$\approx \frac{1}{L}\Big(2\mathrm{Cov}[Y_1, Y_2] + \mathrm{Corr}\left[Y_1 - Y_2,\ F(X_1) - F(X_2)\right]^2\mathrm{Var}[Y_1 - Y_2] - \mathrm{Corr}[Y_1, F(X_1)]^2\mathrm{Var}[Y_1] - \mathrm{Corr}[Y_2, F(X_2)]^2\mathrm{Var}[Y_2]\Big).$$

The results above show that in addition to $\mathrm{Cov}[Y_1, Y_2]$, $\mathrm{Corr}[Y_1 - Y_2, F(X_1) - F(X_2)]$, $\mathrm{Corr}[Y_1, F(X_1)]$, and $\mathrm{Corr}[Y_2, F(X_2)]$ also have a critical impact.

In §6.4, we observe that the unpaired design has larger estimated variance than the paired design, but the effect in prediction-powered evaluation is weaker than in the human-only cases. The correlation diagnostics in Table 4 explain why: the automatic metric is usually less correlated with the human score differences between two systems $Y_1 - Y_2$ than with the individual human scores Y_1 or Y_2. This implies that in a paired design, the power of automatic metrics is utilized less compared to an unpaired design. This suggests a direction for improving automatic metrics.

D Derivation of the Savings Ratio

Starting from

$$\mathrm{Var}\left[\widehat{\delta_{1,2}^{H}}\right] = \frac{1}{L}\mathrm{Var}[Y_1 - Y_2],$$

and

$$\mathrm{Var}\left[\widehat{\delta_{1,2}^{PP}}\right] = \frac{1}{L}\mathrm{Var}[Y_1 - Y_2] - \frac{U\left(\mathrm{Cov}\left[Y_1 - Y_2,\, F_k(X_1) - F_k(X_2)\right]\right)^2}{L(L+U)\mathrm{Var}\left[F_k(X_1) - F_k(X_2)\right]},$$

we have

$$\mathrm{Var}\left[\widehat{\delta_{1,2}^{H}}\right] - \mathrm{Var}\left[\widehat{\delta_{1,2}^{PP}}\right] = \frac{U\left(\mathrm{Cov}\left[Y_1 - Y_2,\, F_k(X_1) - F_k(X_2)\right]\right)^2}{L(L+U)\mathrm{Var}\left[F_k(X_1) - F_k(X_2)\right]}.$$

Therefore,

$$\begin{aligned}
&\frac{\mathrm{Var}\left[\widehat{\delta_{1,2}^{H}}\right] - \mathrm{Var}\left[\widehat{\delta_{1,2}^{PP}}\right]}{\mathrm{Var}\left[\widehat{\delta_{1,2}^{H}}\right]} \\
&= \frac{\frac{U\left(\mathrm{Cov}\left[Y_1 - Y_2,\, F_k(X_1) - F_k(X_2)\right]\right)^2}{L(L+U)\mathrm{Var}\left[F_k(X_1) - F_k(X_2)\right]}}{\frac{1}{L}\mathrm{Var}[Y_1 - Y_2]} \\
&= \frac{U}{L+U} \cdot \frac{\left(\mathrm{Cov}\left[Y_1 - Y_2,\, F_k(X_1) - F_k(X_2)\right]\right)^2}{\mathrm{Var}[Y_1 - Y_2]\,\mathrm{Var}\left[F_k(X_1) - F_k(X_2)\right]} \\
&= \frac{U}{L+U}\mathrm{Corr}\left[Y_1 - Y_2,\, F_k(X_1) - F_k(X_2)\right]^2.
\end{aligned}$$

When $U \gg L$, we have $\frac{U}{L+U} \approx 1$. Hence,

$$\frac{\mathrm{Var}\left[\widehat{\delta_{1,2}^{H}}\right] - \mathrm{Var}\left[\widehat{\delta_{1,2}^{PP}}\right]}{\mathrm{Var}\left[\widehat{\delta_{1,2}^{H}}\right]} \approx \mathrm{Corr}\left[Y_1 - Y_2,\, F_k(X_1) - F_k(X_2)\right]^2.$$

Let L_{eq} denote the number of labeled examples that a human-only evaluation would need in order to match the variance of prediction-powered evaluation with L labeled examples. Then

$$\frac{1}{L_{\mathrm{eq}}}\mathrm{Var}[Y_1 - Y_2] = \mathrm{Var}\left[\widehat{\delta_{1,2}^{PP}}\right].$$

Using

$$\mathrm{Var}\left[\widehat{\delta_{1,2}^{H}}\right] = \frac{1}{L}\mathrm{Var}[Y_1 - Y_2],$$

we obtain

$$\frac{\mathrm{Var}\left[\widehat{\delta_{1,2}^{H}}\right] - \mathrm{Var}\left[\widehat{\delta_{1,2}^{PP}}\right]}{\mathrm{Var}\left[\widehat{\delta_{1,2}^{H}}\right]} = \frac{L_{\mathrm{eq}} - L}{L_{\mathrm{eq}}}.$$

Thus

$$\frac{L_{\mathrm{eq}} - L}{L_{\mathrm{eq}}} \approx \mathrm{Corr}\left[Y_1 - Y_2,\, F_k(X_1) - F_k(X_2)\right]^2.$$

The left-hand side is exactly the proportion of human judgment saved by using prediction-powered evaluation with F_k relative to human-only evaluation, while achieving the same variance.

E Additional Algorithms and Figures

Algorithm 1 Human-only Paired Z-Test

1: **Input:** human scores $\{Y_{1j}, Y_{2j}\}_{j=1}^{L}$
2: **Output:** one-sided p-value for testing $H_0 : \delta_{1,2} \leq 0$ vs. $H_1 : \delta_{1,2} > 0$
3: Compute

$$\widehat{\delta_{1,2}^{H}} \leftarrow \frac{1}{L}\sum_{j=1}^{L}(Y_1 - Y_2)$$

4: Compute

$$\widehat{\mathrm{Var}}[Y_1 - Y_2] \leftarrow \frac{1}{L-1}\sum_{j=1}^{L}(Y_1 - Y_2 - \widehat{\delta_{1,2}^{H}})^2$$

5: Compute

$$\widehat{\mathrm{Var}}[\widehat{\delta_{1,2}^{H}}] \leftarrow \frac{1}{L}\widehat{\mathrm{Var}}[Y_1 - Y_2]$$

6: Compute

$$p_{1,2}^{H} \leftarrow 1 - \Phi\left(\frac{\widehat{\delta_{1,2}^{H}}}{\sqrt{\widehat{\mathrm{Var}}[\widehat{\delta_{1,2}^{H}}]}}\right)$$

7: **return** $p_{1,2}^{H}$

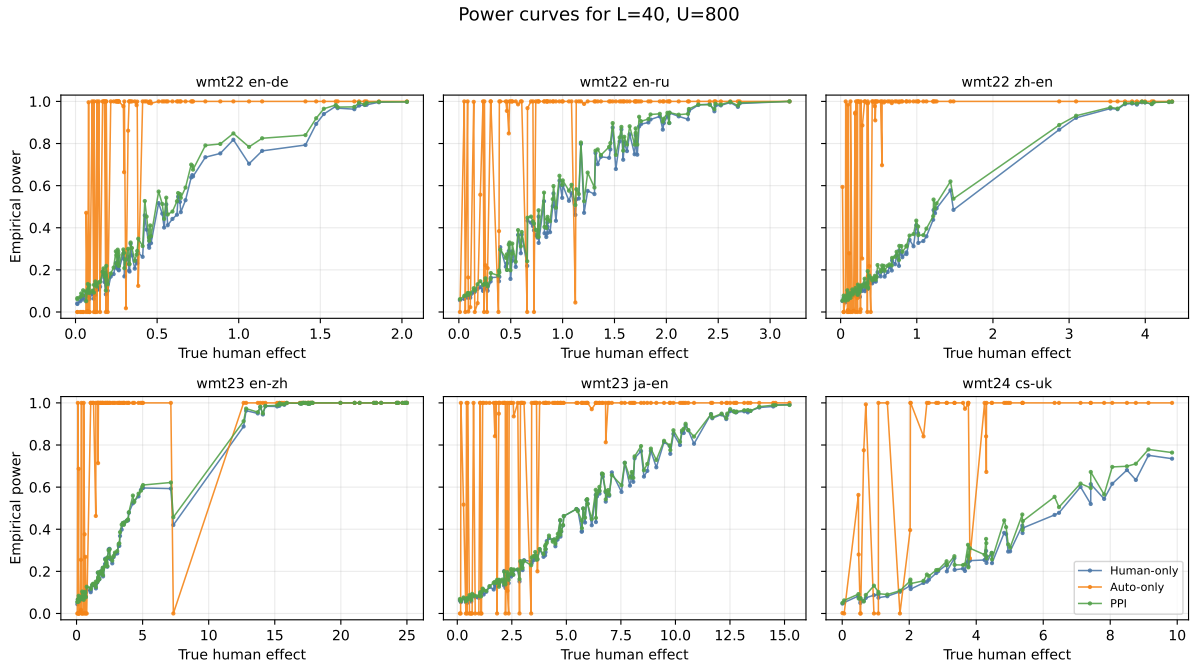

Figure 5: Empirical power curves for $L = 40$ and $U = 800$. Each point is one ordered system pair, and the x-axis is the full-population human-score difference. MetricX is used as the automatic metric. Back to main text

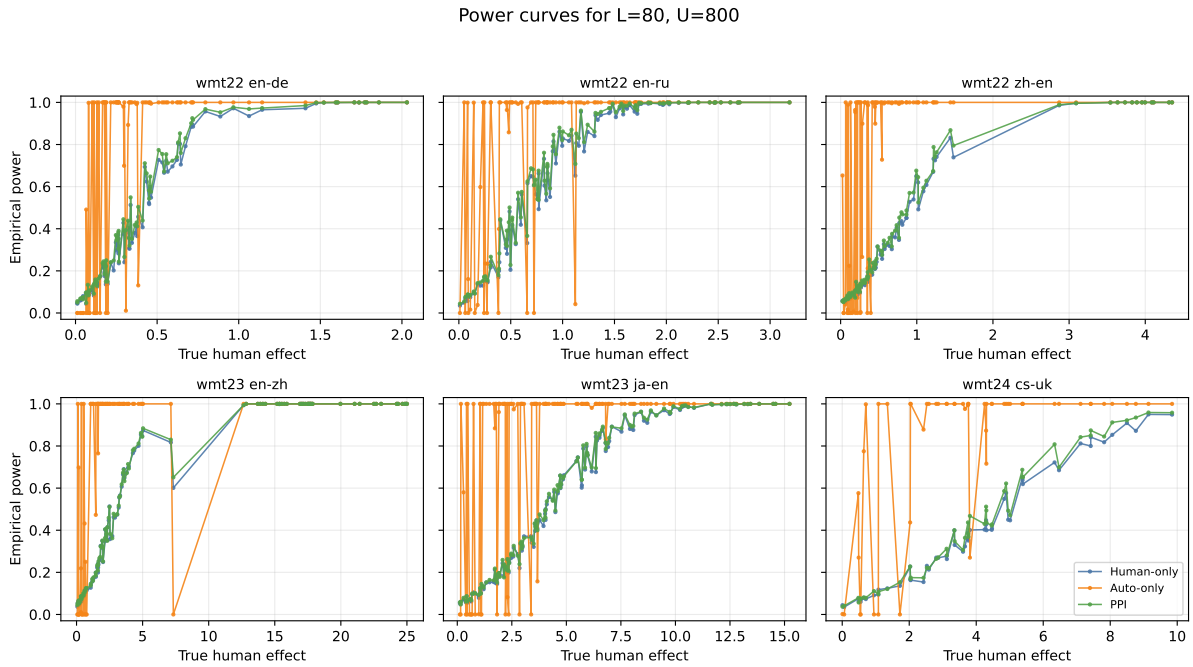

Figure 6: Empirical power curves for $L = 80$ and $U = 800$. Increasing the labeled sample size improves both human-only and prediction-powered tests. MetricX is used as the automatic metric.

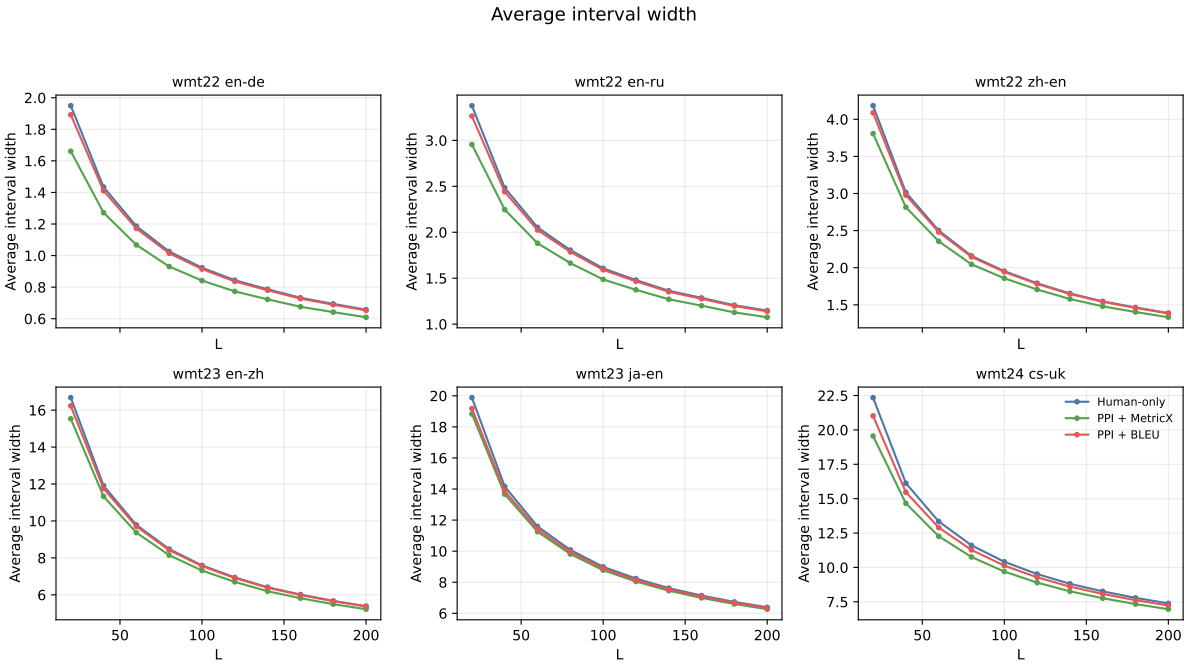

Figure 7: Average two-sided 95% confidence-interval width with $U = 800$. Results are averaged across system pairs.

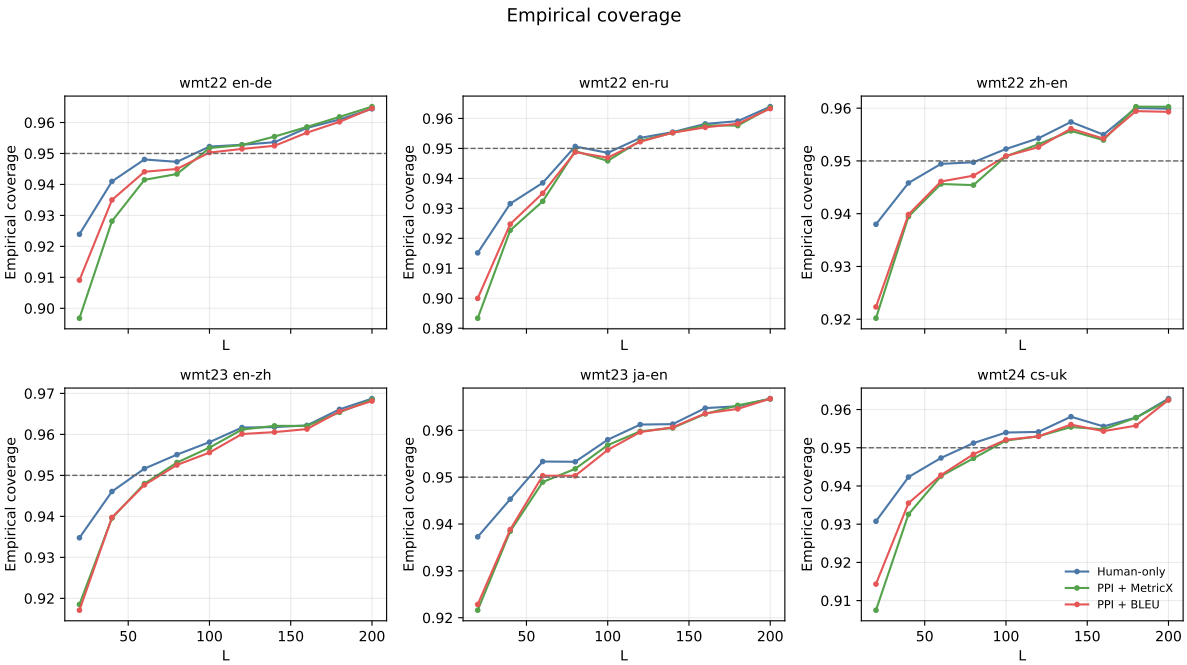

Figure 8: Empirical coverage of the two-sided 95% confidence intervals. The dashed line marks the nominal 95% level.

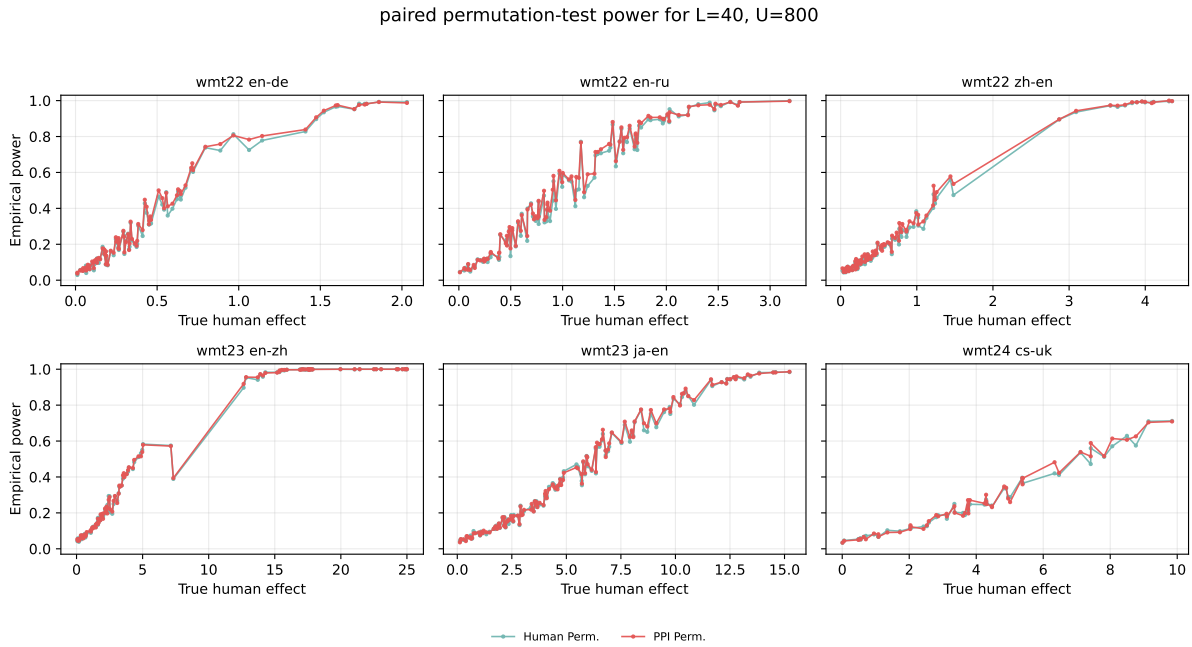

Figure 9: Empirical power curves for $L = 40$ and $U = 800$. "Human Perm" means the human-only paired permutation test and "PPI Perm" means the perdiction-powered paired permutation test. Each point is one ordered system pair, and the x-axis is the full-population human-score difference. MetricX is used as the automatic metric.

Algorithm 2 Auto-only Paired Z-Test

1: **Input:** metric scores $\{F(X_{1j}), F(X_{2j})\}_{j=1}^{M}$
2: **Output:** metric-based one-sided p-value used as a proxy for testing $H_0 : \delta_{1,2} \leq 0$ vs. $H_1 : \delta_{1,2} > 0$
3: Compute

$$\widehat{\delta_{1,2}^{A}} \leftarrow \frac{1}{M}\sum_{j=1}^{M}(F(X_{1j}) - F(X_{2j}))$$

4: Compute

$$\widehat{\mathrm{Var}}[F(X_{1j}) - F(X_{2j})] \leftarrow \tfrac{1}{M-1}\textstyle\sum_{j=1}^{M}(F(X_{1j}) - F(X_{2j}) - \widehat{\delta_{1,2}^{A}})^2$$

5: Compute

$$\widehat{\mathrm{Var}}[\widehat{\delta_{1,2}^{A}}] \leftarrow \frac{1}{M}\widehat{\mathrm{Var}}[F(X_{1j}) - F(X_{2j})]$$

6: Compute

$$p_{1,2}^{A} \leftarrow 1 - \Phi\Big(\frac{\widehat{\delta_{1,2}^{A}}}{\sqrt{\widehat{\mathrm{Var}}[\widehat{\delta_{1,2}^{A}}]}}\Big)$$

7: **return** $p_{1,2}^{A}$

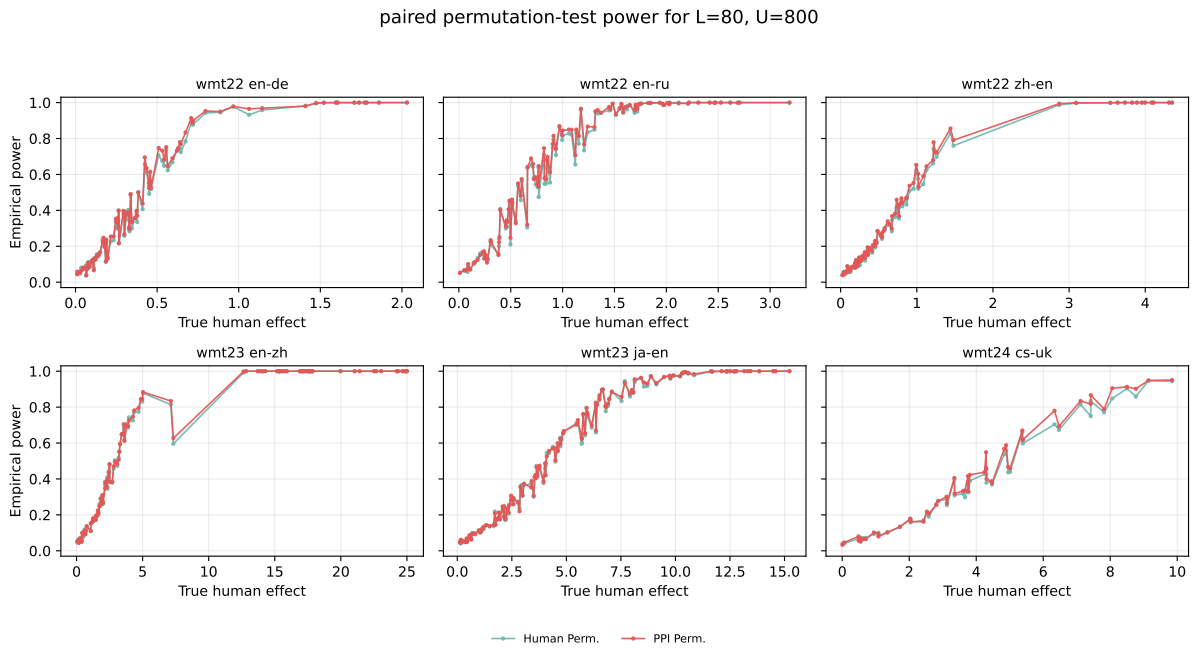

Figure 10: Empirical power curves for $L = 80$ and $U = 800$. "Human Perm" means the human-only paired permutation test and "PPI Perm" means the perdiction-powered paired permutation test. Each point is one ordered system pair, and the x-axis is the full-population human-score difference. MetricX is used as the automatic metric.

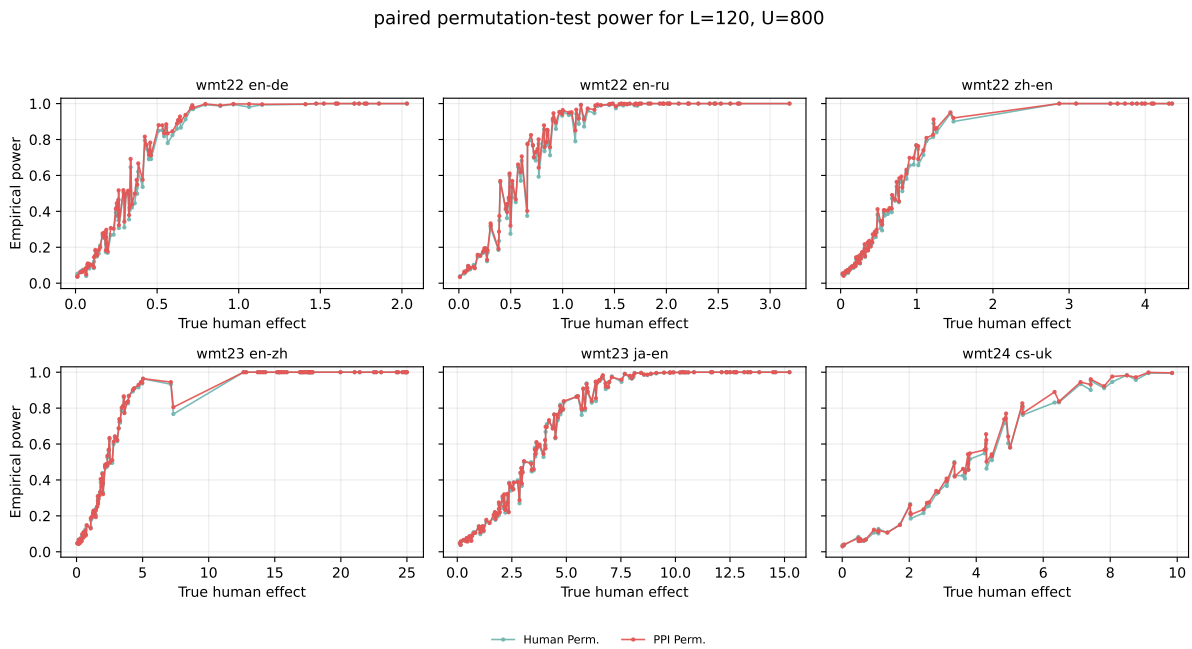

Figure 11: Empirical power curves for $L = 120$ and $U = 800$. "Human Perm" means the human-only paired permutation test and "PPI Perm" means the perdiction-powered paired permutation test. Each point is one ordered system pair, and the x-axis is the full-population human-score difference. MetricX is used as the automatic metric.

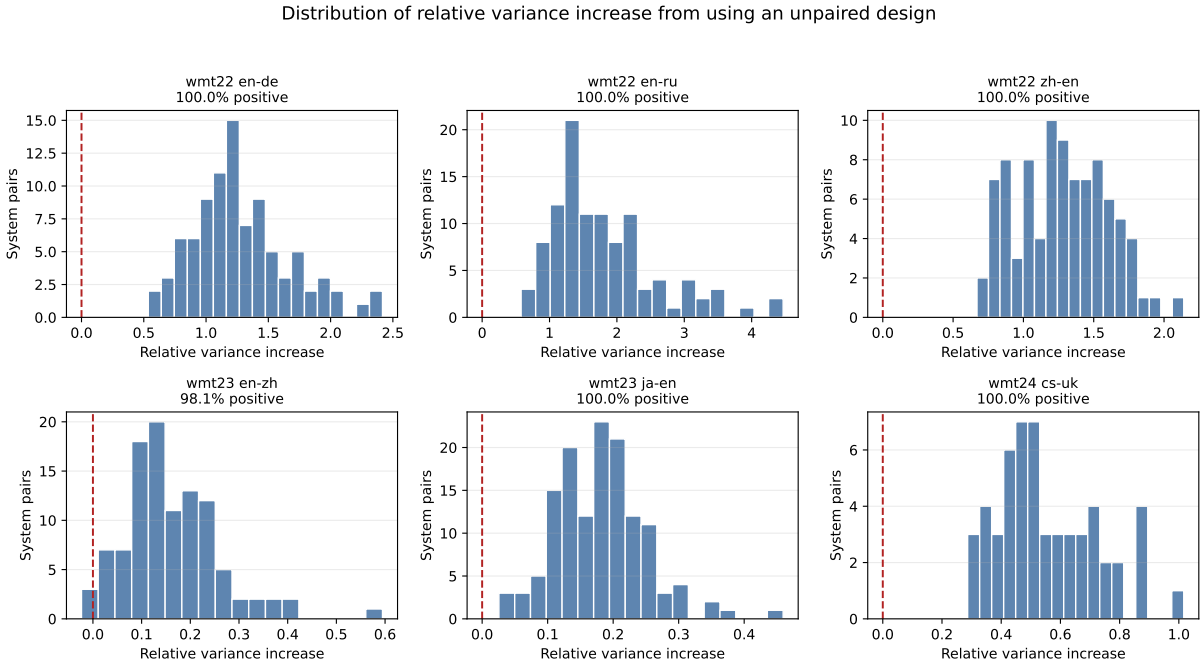

Figure 12: Distribution of the relative variance increase from using an unpaired human-only design instead of the paired design. The dashed vertical line marks zero. Values to the right indicate larger variance under the unpaired design.

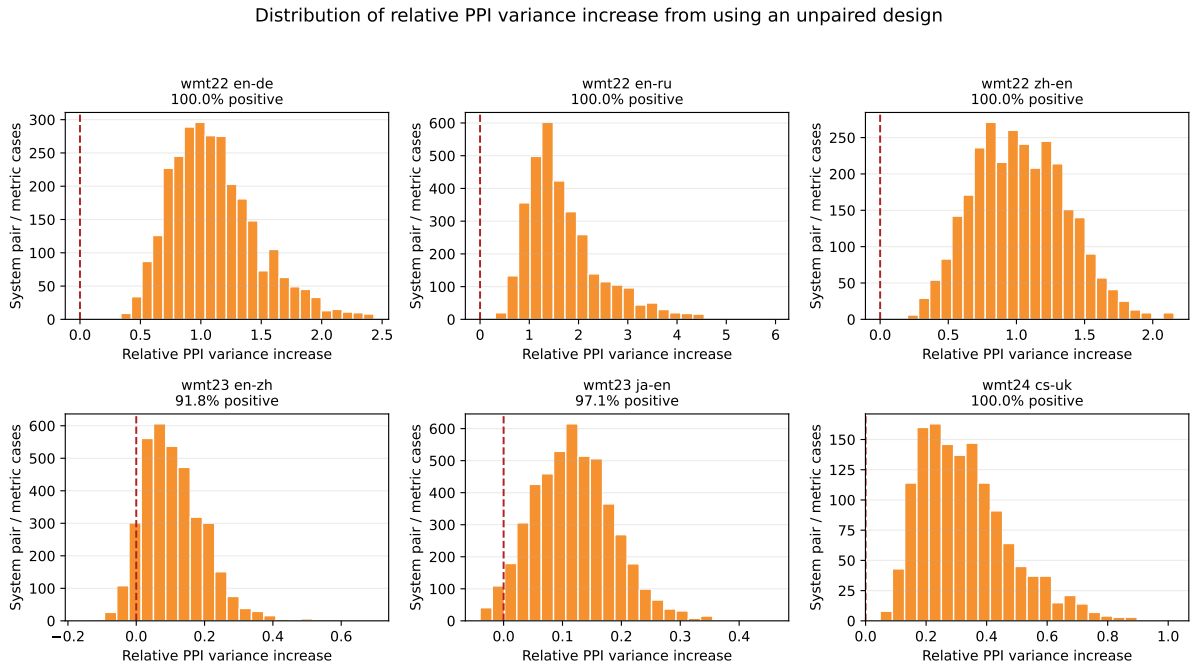

Figure 13: Distribution of the relative prediction-powered variance increase from using an unpaired design instead of the paired design. Each observation is one system-pair/metric case. The dashed vertical line marks zero. Values to the right indicate larger variance under the unpaired design.

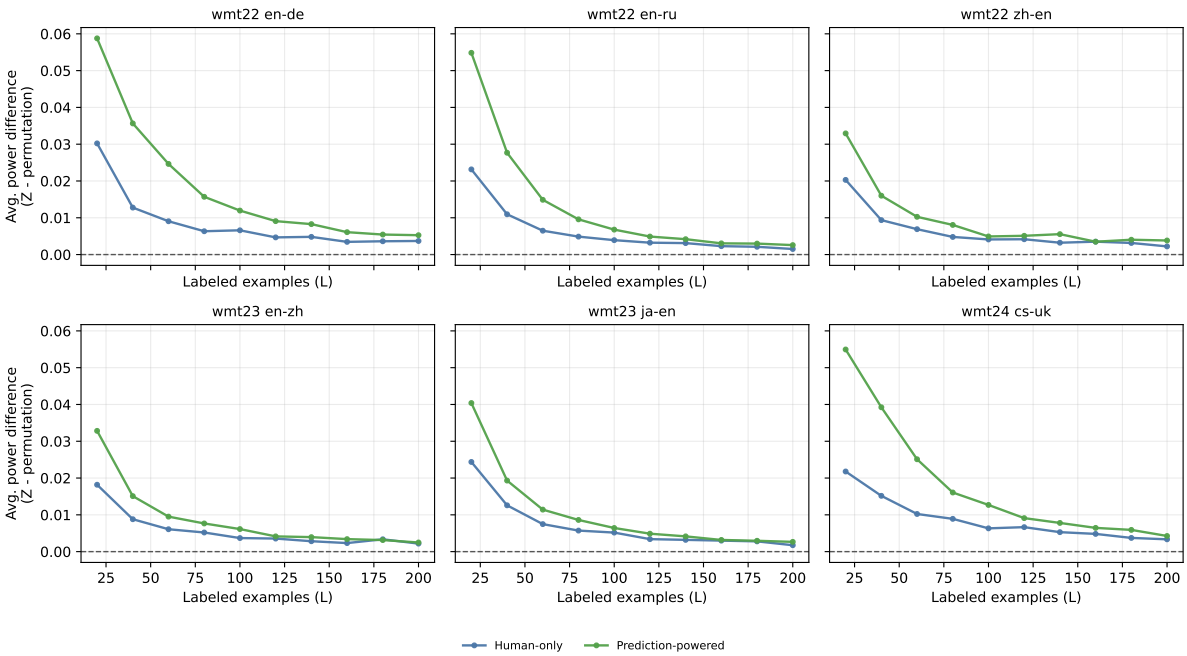

Figure 14: Average power difference across all system pairs from using the paired permutation test instead of the paired Z-test. Positive values indicate higher empirical power for the Z-test.

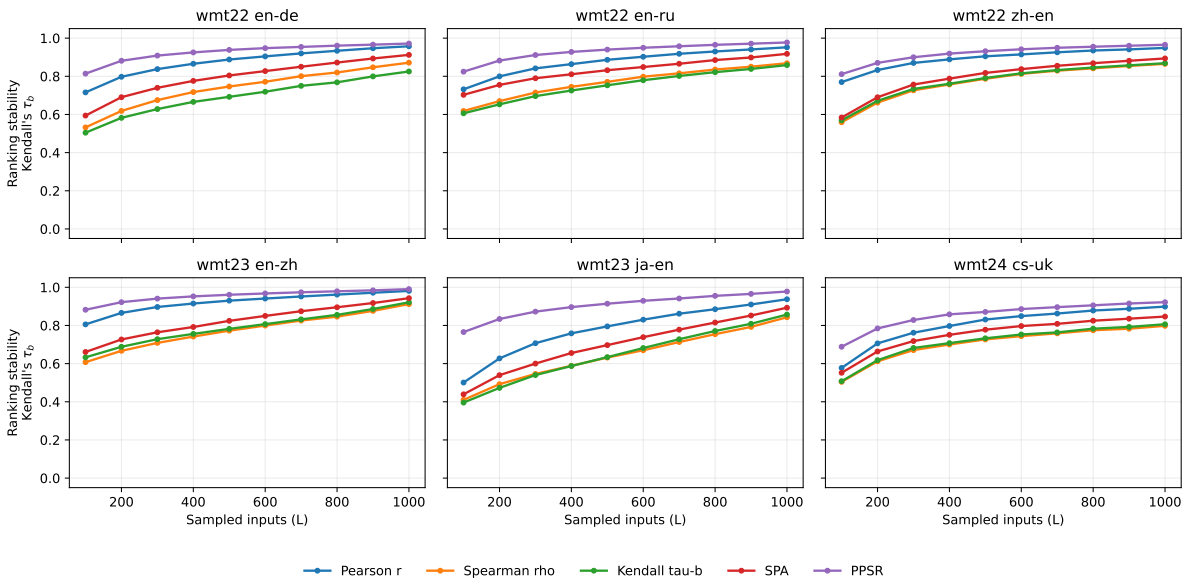

Figure 15: Ranking stability of system-level meta-metrics. Higher values indicate more stable metric rankings. Note that for meta-evaluation, only labeled examples are used and the unlabeled examples are unused.

Metric	r	ρ	τ_b	SPA	PPSR
metricx_xl_MQM_2020-refA	0.939 (5)	0.771 (7)	0.582 (8)	0.819 (6)	0.155 (1)
metricx_xxl_MQM_2020-refA	0.953 (4)	0.837 (2)	0.626 (2)	0.831 (3)	0.155 (2)
COMET-22-refA	0.922 (6)	0.749 (11)	0.604 (6)	0.810 (7)	0.140 (3)
metricx_xl_DA_2019-refA	0.969 (2)	0.829 (4)	0.582 (8)	0.829 (4)	0.138 (4)
metricx_xxl_DA_2019-refA	0.972 (1)	0.842 (1)	0.626 (2)	0.835 (2)	0.136 (5)
UniTE-refA	0.875 (11)	0.763 (8)	0.538 (12)	0.785 (11)	0.116 (6)
UniTE-ref-refA	0.866 (14)	0.763 (8)	0.538 (12)	0.784 (12)	0.112 (7)
BLEURT-20-refA	0.895 (8)	0.815 (5)	0.604 (6)	0.821 (5)	0.102 (8)
COMETKiwi-src	0.885 (10)	0.705 (17)	0.560 (10)	0.785 (10)	0.101 (9)
COMET-20-refA	0.964 (3)	0.833 (3)	0.670 (1)	0.848 (1)	0.091 (10)
Cross-QE-src	0.895 (9)	0.710 (15)	0.516 (16)	0.775 (14)	0.085 (11)
UniTE-src-src	0.797 (21)	0.631 (24)	0.451 (25)	0.737 (23)	0.084 (12)
MS-COMET-22-refA	0.875 (12)	0.758 (10)	0.626 (2)	0.804 (8)	0.068 (13)
COMET-QE-src	0.676 (28)	0.662 (22)	0.495 (21)	0.761 (18)	0.066 (14)
SEScore-refA	0.855 (15)	0.675 (20)	0.495 (21)	0.766 (16)	0.062 (15)
MATESE-refA	0.872 (13)	0.749 (11)	0.516 (16)	0.778 (13)	0.051 (16)
MS-COMET-QE-22-src	0.748 (24)	0.675 (20)	0.516 (16)	0.740 (22)	0.050 (17)
YiSi-1-refA	0.897 (7)	0.815 (5)	0.626 (2)	0.800 (9)	0.038 (18)
MEE4-refA	0.833 (17)	0.719 (13)	0.538 (12)	0.770 (15)	0.034 (19)
BERTScore-refA	0.818 (19)	0.679 (18)	0.516 (16)	0.757 (20)	0.033 (20)
MATESE-QE-src	0.752 (22)	0.560 (28)	0.341 (30)	0.681 (30)	0.032 (21)
MEE2-refA	0.824 (18)	0.719 (13)	0.538 (12)	0.758 (19)	0.027 (22)
chrF-refA	0.848 (16)	0.710 (15)	0.560 (10)	0.763 (17)	0.027 (23)
HWTSC-Teacher-Sim-src	0.641 (29)	0.565 (27)	0.429 (27)	0.702 (28)	0.022 (24)
MEE-refA	0.816 (20)	0.679 (18)	0.516 (16)	0.743 (21)	0.020 (25)
f200spBLEU-refA	0.723 (26)	0.618 (25)	0.473 (24)	0.717 (25)	0.020 (26)
f101spBLEU-refA	0.727 (25)	0.662 (22)	0.495 (21)	0.721 (24)	0.019 (27)
KG-BERTScore-src	0.614 (30)	0.481 (30)	0.363 (29)	0.694 (29)	0.017 (28)
HWTSC-TLM-src	0.749 (23)	0.556 (29)	0.429 (27)	0.713 (26)	0.014 (29)
BLEU-refA	0.685 (27)	0.609 (26)	0.451 (25)	0.708 (27)	0.013 (30)
REUSE-src	-0.651 (31)	-0.490 (31)	-0.363 (31)	0.327 (31)	0.008 (31)

Table 5: The scores (and ranks) of automatic metric under each system-level meta-metric for WMT22 en-de. Metrics are sorted by PPSR. Tied scores receive the same rank.

Metric	r	ρ	τ_b	SPA	PPSR
mqm.unb	0.985 (1)	0.950 (1)	0.867 (1)	0.941 (1)	0.469 (1)
metricx_xxl_MQM_2020-refA	0.953 (5)	0.921 (4)	0.829 (4)	0.915 (4)	0.135 (2)
metricx_xl_MQM_2020-refA	0.932 (7)	0.904 (7)	0.790 (7)	0.900 (6)	0.128 (3)
COMET-22-refA	0.908 (8)	0.875 (13)	0.714 (13)	0.880 (8)	0.102 (4)
UniTE-refA	0.894 (9)	0.886 (8)	0.752 (8)	0.878 (9)	0.096 (5)
metricx_xxl_DA_2019-refA	0.979 (2)	0.943 (3)	0.848 (2)	0.933 (2)	0.094 (6)
UniTE-ref-refA	0.841 (12)	0.854 (14)	0.752 (8)	0.873 (10)	0.091 (7)
metricx_xl_DA_2019-refA	0.975 (3)	0.946 (2)	0.848 (2)	0.923 (3)	0.087 (8)
UniTE-src-src	0.772 (23)	0.789 (23)	0.657 (20)	0.825 (21)	0.070 (9)
BLEURT-20-refA	0.956 (4)	0.921 (4)	0.829 (4)	0.903 (5)	0.069 (10)
COMETKiwi-src	0.763 (24)	0.782 (24)	0.619 (25)	0.821 (24)	0.065 (11)
MATESE-refA	0.795 (22)	0.832 (20)	0.714 (13)	0.847 (13)	0.061 (12)
COMET-20-refA	0.940 (6)	0.921 (4)	0.810 (6)	0.893 (7)	0.061 (13)
Cross-QE-src	0.802 (21)	0.814 (22)	0.657 (20)	0.826 (20)	0.060 (14)
MS-COMET-22-refA	0.832 (14)	0.843 (17)	0.695 (18)	0.847 (12)	0.057 (15)
COMET-QE-src	0.483 (30)	0.632 (28)	0.600 (26)	0.799 (26)	0.055 (16)
MS-COMET-QE-22-src	0.718 (25)	0.779 (25)	0.638 (24)	0.812 (25)	0.045 (17)
MATESE-QE-src	0.665 (28)	0.582 (30)	0.467 (28)	0.768 (27)	0.040 (18)
YiSi-1-refA	0.882 (10)	0.879 (12)	0.714 (13)	0.847 (11)	0.033 (19)
BERTScore-refA	0.817 (18)	0.882 (9)	0.733 (10)	0.833 (19)	0.025 (20)
MEE4-refA	0.809 (20)	0.882 (9)	0.733 (10)	0.843 (14)	0.022 (21)
chrF-refA	0.874 (11)	0.836 (19)	0.657 (20)	0.837 (17)	0.020 (22)
MEE2-refA	0.825 (16)	0.882 (9)	0.733 (10)	0.838 (15)	0.019 (23)
f200spBLEU-refA	0.827 (15)	0.854 (14)	0.714 (13)	0.838 (16)	0.018 (24)
f101spBLEU-refA	0.821 (17)	0.854 (14)	0.714 (13)	0.833 (18)	0.018 (25)
MEE-refA	0.835 (13)	0.839 (18)	0.657 (20)	0.823 (23)	0.015 (26)
BLEU-refA	0.813 (19)	0.821 (21)	0.676 (19)	0.824 (22)	0.014 (27)
KG-BERTScore-src	0.716 (26)	0.661 (27)	0.467 (28)	0.759 (28)	0.013 (28)
HWTSC-Teacher-Sim-src	0.703 (27)	0.718 (26)	0.543 (27)	0.747 (29)	0.013 (29)
HWTSC-TLM-src	0.642 (29)	0.632 (28)	0.467 (28)	0.729 (30)	0.011 (30)
REUSE-src	-0.381 (31)	-0.493 (31)	-0.429 (31)	0.286 (31)	0.005 (31)

Table 6: The scores (and ranks) of automatic metric under each system-level meta-metric for WMT22 en-ru. Metrics are sorted by PPSR. Tied scores receive the same rank.

Metric	r	ρ	τ_b	SPA	PPSR
COMET-22-refA	0.942 (5)	0.859 (4)	0.736 (3)	0.876 (3)	0.093 (1)
metricx_xl_MQM_2020-refA	0.914 (10)	0.877 (3)	0.736 (3)	0.873 (4)	0.091 (2)
metricx_xxl_MQM_2020-refA	0.920 (9)	0.842 (7)	0.692 (7)	0.862 (5)	0.087 (3)
metricx_xl_DA_2019-refA	0.982 (2)	0.895 (2)	0.758 (1)	0.895 (1)	0.070 (4)
UniTE-refA	0.914 (11)	0.837 (8)	0.692 (7)	0.849 (10)	0.069 (5)
COMETKiwi-src	0.866 (19)	0.785 (12)	0.648 (12)	0.828 (12)	0.067 (6)
UniTE-ref-refA	0.892 (15)	0.820 (11)	0.670 (11)	0.839 (11)	0.065 (7)
metricx_xxl_DA_2019-refA	0.984 (1)	0.908 (1)	0.758 (1)	0.894 (2)	0.064 (8)
Cross-QE-src	0.870 (18)	0.736 (16)	0.582 (16)	0.799 (16)	0.057 (9)
MATESE-refA	0.856 (20)	0.855 (5)	0.714 (6)	0.858 (6)	0.055 (10)
UniTE-src-src	0.874 (16)	0.635 (17)	0.516 (17)	0.775 (17)	0.053 (11)
BLEURT-20-refA	0.938 (6)	0.837 (8)	0.692 (7)	0.853 (7)	0.051 (12)
SEScore-refA	0.944 (4)	0.754 (13)	0.604 (15)	0.815 (15)	0.045 (13)
MATESE-QE-src	0.767 (23)	0.837 (8)	0.692 (7)	0.850 (9)	0.039 (14)
COMET-QE-src	0.569 (27)	0.749 (15)	0.626 (13)	0.816 (14)	0.039 (15)
COMET-20-refA	0.970 (3)	0.754 (13)	0.626 (13)	0.827 (13)	0.037 (16)
BERTScore-refA	0.924 (8)	0.578 (20)	0.451 (19)	0.747 (19)	0.032 (17)
YiSi-1-refA	0.935 (7)	0.591 (19)	0.451 (19)	0.737 (20)	0.030 (18)
MS-COMET-22-refA	0.909 (12)	0.846 (6)	0.736 (3)	0.852 (8)	0.024 (19)
MS-COMET-QE-22-src	0.897 (14)	0.604 (18)	0.473 (18)	0.769 (18)	0.022 (20)
MEE4-refA	0.905 (13)	0.569 (21)	0.451 (19)	0.723 (21)	0.019 (21)
MEE2-refA	0.872 (17)	0.446 (25)	0.363 (23)	0.684 (23)	0.014 (22)
HWTSC-Teacher-Sim-src	0.356 (30)	0.499 (22)	0.363 (23)	0.678 (24)	0.011 (23)
chrF-refA	0.812 (22)	0.429 (28)	0.297 (28)	0.660 (28)	0.010 (24)
KG-BERTScore-src	0.553 (28)	0.486 (23)	0.385 (22)	0.696 (22)	0.009 (25)
MEE-refA	0.824 (21)	0.349 (29)	0.297 (28)	0.650 (29)	0.008 (26)
f200spBLEU-refA	0.728 (24)	0.437 (26)	0.341 (26)	0.668 (26)	0.007 (27)
f101spBLEU-refA	0.718 (25)	0.437 (26)	0.341 (26)	0.665 (27)	0.007 (28)
HWTSC-TLM-src	0.460 (29)	0.455 (24)	0.363 (23)	0.669 (25)	0.006 (29)
BLEU-refA	0.658 (26)	0.345 (30)	0.275 (30)	0.646 (30)	0.005 (30)
REUSE-src	-0.142 (31)	-0.332 (31)	-0.231 (31)	0.386 (31)	0.001 (31)

Table 7: The scores (and ranks) of automatic metric under each system-level meta-metric for WMT22 zh-en. Metrics are sorted by PPSR. Tied scores receive the same rank.

Metric	r	ρ	τ_b	SPA	PPSR
CometKiwi-src	0.992 (5)	0.950 (4)	0.867 (2)	0.925 (4)	0.186 (1)
KG-BERTScore-src	0.992 (4)	0.950 (4)	0.867 (2)	0.925 (5)	0.186 (2)
CometKiwi-XL-src	0.985 (10)	0.921 (10)	0.810 (9)	0.916 (11)	0.160 (3)
MS-COMET-QE-22-src	0.993 (3)	0.950 (4)	0.848 (5)	0.930 (2)	0.157 (4)
COMET-refA	0.995 (2)	0.954 (2)	0.867 (2)	0.920 (7)	0.149 (5)
cometoid22-wmt23-src	0.996 (1)	0.971 (1)	0.886 (1)	0.946 (1)	0.148 (6)
CometKiwi-XXL-src	0.980 (12)	0.925 (9)	0.829 (7)	0.917 (10)	0.145 (7)
BLEURT-20-refA	0.986 (9)	0.875 (22)	0.771 (14)	0.869 (22)	0.140 (8)
cometoid22-wmt22-src	0.991 (6)	0.921 (10)	0.810 (9)	0.922 (6)	0.126 (9)
XLsim-refA	0.981 (11)	0.807 (23)	0.695 (23)	0.826 (23)	0.124 (10)
YiSi-1-refA	0.972 (14)	0.696 (25)	0.562 (25)	0.773 (25)	0.123 (11)
cometoid22-wmt21-src	0.990 (7)	0.932 (8)	0.810 (9)	0.919 (9)	0.121 (12)
prismRef-refA	0.970 (15)	0.725 (24)	0.619 (24)	0.795 (24)	0.118 (13)
MetricX-23-c-refA	0.936 (22)	0.954 (2)	0.829 (7)	0.927 (3)	0.115 (14)
BERTscore-refA	0.972 (13)	0.586 (29)	0.486 (28)	0.744 (27)	0.113 (15)
XCOMET-Ensemble-refA	0.944 (20)	0.911 (17)	0.771 (14)	0.899 (18)	0.110 (16)
GEMBA-MQM-src	0.967 (16)	0.946 (7)	0.848 (5)	0.919 (8)	0.090 (17)
XCOMET-QE-Ensemble-src	0.908 (29)	0.918 (12)	0.790 (12)	0.905 (14)	0.085 (18)
mre-score-labse-regular-refA	0.989 (8)	0.896 (20)	0.733 (21)	0.877 (20)	0.085 (19)
MetricX-23-QE-b-src	0.967 (17)	0.904 (18)	0.771 (14)	0.910 (12)	0.076 (20)
MetricX-23-QE-c-src	0.909 (28)	0.914 (13)	0.771 (14)	0.900 (17)	0.075 (21)
prismSrc-src	0.931 (23)	0.168 (34)	0.105 (34)	0.549 (34)	0.074 (22)
MetricX-23-QE-src	0.950 (18)	0.914 (13)	0.790 (12)	0.904 (15)	0.066 (23)
MetricX-23-b-refA	0.930 (24)	0.914 (13)	0.771 (14)	0.900 (16)	0.065 (24)
MetricX-23-refA	0.892 (30)	0.914 (13)	0.771 (14)	0.908 (13)	0.061 (25)
XCOMET-XXL-refA	0.886 (31)	0.904 (18)	0.752 (20)	0.891 (19)	0.053 (26)
XCOMET-XL-refA	0.796 (32)	0.893 (21)	0.733 (21)	0.876 (21)	0.044 (27)
tokengram_F-refA	0.943 (21)	0.614 (27)	0.505 (27)	0.740 (28)	0.042 (28)
f200spBLEU-refA	0.912 (26)	0.532 (31)	0.429 (30)	0.705 (31)	0.039 (29)
chrF-refA	0.945 (19)	0.596 (28)	0.467 (29)	0.726 (29)	0.037 (30)
embed_llama-refA	0.924 (25)	0.514 (32)	0.410 (32)	0.708 (30)	0.030 (31)
eBLEU-refA	0.910 (27)	0.554 (30)	0.429 (30)	0.704 (32)	0.011 (32)
BLEU-refA	0.764 (33)	0.671 (26)	0.524 (26)	0.753 (26)	0.003 (33)
Random-sysname-src	0.045 (34)	0.336 (33)	0.200 (33)	0.598 (33)	0.001 (34)

Table 8: The scores (and ranks) of automatic metric under each system-level meta-metric for WMT23 en-zh. Metrics are sorted by PPSR. Tied scores receive the same rank.

Metric	r	ρ	τ_b	SPA	PPSR
CometKiwi-XXL-src	0.987 (2)	0.975 (1)	0.912 (1)	0.948 (2)	0.101 (1)
COMET-refA	0.967 (13)	0.953 (10)	0.853 (8)	0.923 (14)	0.096 (2)
BLEURT-20-refA	0.959 (21)	0.953 (10)	0.853 (8)	0.920 (17)	0.093 (3)
KG-BERTScore-src	0.973 (8)	0.973 (3)	0.897 (4)	0.938 (6)	0.091 (4)
CometKiwi-src	0.973 (9)	0.973 (3)	0.897 (4)	0.938 (7)	0.091 (5)
MetricX-23-c-refA	0.962 (18)	0.961 (7)	0.868 (6)	0.939 (5)	0.087 (6)
cometoid22-wmt23-src	0.967 (14)	0.958 (8)	0.853 (8)	0.924 (11)	0.086 (7)
CometKiwi-XL-src	0.986 (3)	0.975 (1)	0.912 (1)	0.949 (1)	0.085 (8)
XCOMET-Ensemble-refA	0.951 (22)	0.949 (14)	0.838 (15)	0.922 (16)	0.081 (9)
cometoid22-wmt22-src	0.944 (25)	0.944 (19)	0.824 (19)	0.912 (20)	0.071 (10)
cometoid22-wmt21-src	0.947 (23)	0.941 (20)	0.809 (25)	0.913 (19)	0.069 (11)
YiSi-1-refA	0.974 (7)	0.941 (20)	0.838 (15)	0.924 (13)	0.066 (12)
MS-COMET-QE-22-src	0.905 (33)	0.949 (14)	0.838 (15)	0.909 (24)	0.065 (13)
prismRef-refA	0.970 (12)	0.931 (25)	0.794 (27)	0.902 (30)	0.064 (14)
MetricX-23-QE-c-src	0.966 (15)	0.953 (10)	0.853 (8)	0.932 (8)	0.062 (15)
GEMBA-MQM-src	0.985 (4)	0.963 (6)	0.868 (6)	0.945 (3)	0.062 (16)
MetricX-23-QE-b-src	0.977 (6)	0.949 (14)	0.853 (8)	0.929 (9)	0.058 (17)
BERTscore-refA	0.971 (11)	0.951 (13)	0.853 (8)	0.924 (12)	0.057 (18)
XLsim-refA	0.987 (1)	0.968 (5)	0.912 (1)	0.939 (4)	0.056 (19)
XCOMET-XXL-refA	0.942 (26)	0.934 (24)	0.794 (27)	0.906 (28)	0.054 (20)
XCOMET-QE-Ensemble-src	0.944 (24)	0.949 (14)	0.838 (15)	0.925 (10)	0.054 (21)
MetricX-23-b-refA	0.940 (29)	0.931 (25)	0.779 (31)	0.902 (29)	0.048 (22)
MetricX-23-QE-src	0.941 (28)	0.939 (22)	0.809 (25)	0.910 (22)	0.045 (23)
chrF-refA	0.965 (16)	0.931 (25)	0.824 (19)	0.912 (21)	0.044 (24)
XCOMET-XL-refA	0.926 (30)	0.946 (18)	0.824 (19)	0.908 (25)	0.043 (25)
tokengram_F-refA	0.965 (17)	0.931 (25)	0.824 (19)	0.913 (18)	0.042 (26)
MetricX-23-refA	0.918 (31)	0.919 (31)	0.765 (32)	0.892 (31)	0.042 (27)
mre-score-labse-regular-refA	0.981 (5)	0.956 (9)	0.853 (8)	0.922 (15)	0.032 (28)
eBLEU-refA	0.961 (20)	0.936 (23)	0.824 (19)	0.909 (23)	0.028 (29)
f200spBLEU-refA	0.962 (19)	0.926 (29)	0.824 (19)	0.906 (27)	0.021 (30)
embed_llama-refA	0.971 (10)	0.926 (29)	0.794 (27)	0.908 (26)	0.016 (31)
BLEU-refA	0.942 (27)	0.919 (31)	0.794 (27)	0.891 (32)	0.016 (32)
MaTESe-refA	0.911 (32)	0.895 (33)	0.735 (33)	0.884 (33)	0.013 (33)
prismSrc-src	-0.773 (35)	-0.706 (35)	-0.559 (35)	0.240 (35)	0.006 (34)
Random-sysname-src	0.288 (34)	0.279 (34)	0.191 (34)	0.605 (34)	0.001 (35)

Table 9: The scores (and ranks) of automatic metric under each system-level meta-metric for WMT23 ja-en. Metrics are sorted by PPSR. Tied scores receive the same rank.

Metric	r	ρ	τ_b	SPA	PPSR
MetricX-24-refA	0.924 (6)	0.900 (5)	0.745 (8)	0.890 (11)	0.124 (1)
PrismRefMedium-refA	0.949 (2)	0.909 (4)	0.782 (4)	0.926 (2)	0.114 (2)
COMET-22-refA	0.921 (8)	0.973 (1)	0.891 (1)	0.917 (3)	0.112 (3)
BLEURT-20-refA	0.981 (1)	0.973 (1)	0.891 (1)	0.960 (1)	0.111 (4)
PrismRefSmall-refA	0.931 (3)	0.900 (5)	0.745 (8)	0.905 (6)	0.110 (5)
MetricX-24-Hybrid-refA	0.898 (11)	0.882 (10)	0.709 (10)	0.883 (12)	0.106 (6)
metametrics_mt_mqm_kendall-refA	0.801 (16)	0.773 (15)	0.600 (15)	0.821 (16)	0.104 (7)
metametrics_mt_mqm_hybrid_kendall-refA	0.804 (15)	0.773 (15)	0.600 (15)	0.821 (15)	0.103 (8)
chrfS-refA	0.908 (9)	0.900 (5)	0.782 (4)	0.896 (8)	0.092 (9)
chrF-refA	0.931 (4)	0.900 (5)	0.782 (4)	0.901 (7)	0.079 (10)
BERTScore-refA	0.819 (13)	0.809 (14)	0.673 (14)	0.826 (14)	0.079 (11)
damonmonli-refA	0.865 (12)	0.855 (11)	0.709 (10)	0.892 (9)	0.077 (12)
YiSi-1-refA	0.925 (5)	0.955 (3)	0.855 (3)	0.914 (4)	0.072 (13)
gemba_esa-src	0.785 (17)	0.745 (18)	0.564 (18)	0.777 (20)	0.069 (14)
XCOMET-refA	0.766 (19)	0.709 (20)	0.527 (19)	0.782 (19)	0.063 (15)
monmonli-refA	0.818 (14)	0.845 (12)	0.709 (10)	0.848 (13)	0.055 (16)
spBLEU-refA	0.922 (7)	0.900 (5)	0.782 (4)	0.912 (5)	0.055 (17)
MetricX-24-Hybrid-QE-src	0.748 (20)	0.736 (19)	0.527 (19)	0.790 (18)	0.045 (18)
MetricX-24-QE-src	0.776 (18)	0.773 (15)	0.600 (15)	0.820 (17)	0.044 (19)
BLEU-refA	0.899 (10)	0.845 (12)	0.709 (10)	0.892 (10)	0.043 (20)
XCOMET-QE-src	0.566 (22)	0.482 (22)	0.345 (22)	0.698 (22)	0.027 (21)
CometKiwi-src	0.469 (23)	0.427 (23)	0.273 (23)	0.683 (23)	0.025 (22)
CometKiwi-XXL-src	0.699 (21)	0.627 (21)	0.455 (21)	0.774 (21)	0.022 (23)
metametrics_mt_mqm_qe_kendall.seg.s-src	0.330 (24)	0.145 (24)	0.127 (24)	0.595 (24)	0.020 (24)
XLsimDA-src	-0.318 (25)	-0.009 (25)	-0.018 (25)	0.475 (25)	0.009 (25)

Table 10: The scores (and ranks) of automatic metric under each system-level meta-metric for WMT24 cs-uk. Metrics are sorted by PPSR. Tied scores receive the same rank.

Algorithm 3 Prediction-powered Paired Z-Test

Input: system outputs and their human scores $\{X_{1j}, X_{2j}, Y_{1j}, Y_{2j}\}_{j=1}^{L}$, system outputs without human scores $\{X_{1j}, X_{2j}\}_{j=L+1}^{L+U}$, automatic metric F.
Output: one-sided p-value for testing $H_0 : \delta_{1,2} \leq 0$ vs. $H_1 : \delta_{1,2} > 0$
Compute $M \leftarrow L + U$
Set $d_j \leftarrow Y_{1j} - Y_{2j}$ for $j = 1, \ldots, L$ and $f_j \leftarrow F(X_{1j}) - F(X_{2j})$ for $j = 1, \ldots, M$
Compute

$$\bar{d}_L \leftarrow \frac{1}{L}\sum_{j=1}^{L} d_j$$

Compute

$$\widehat{\mathrm{Var}}[d] \leftarrow \frac{1}{L-1}\sum_{j=1}^{L}(d_j - \bar{d}_L)^2$$

Compute

$$\bar{f}_L \leftarrow \frac{1}{L}\sum_{j=1}^{L} f_j, \qquad \bar{f}_U \leftarrow \frac{1}{U}\sum_{j=L+1}^{L+U} f_j,$$
$$\bar{f}_M \leftarrow \frac{1}{M}\sum_{j=1}^{M} f_j$$

Compute

$$\widehat{\mathrm{Var}}[f] \leftarrow \frac{1}{M-1}\sum_{j=1}^{M}(f_j - \bar{f}_M)^2$$

Compute

$$\widehat{\mathrm{Cov}}[d, f] \leftarrow \frac{1}{L-1}\sum_{j=1}^{L}(d_j - \bar{d}_L)(f_j - \bar{f}_L)$$

Compute

$$\lambda^\star \leftarrow \frac{\widehat{\mathrm{Cov}}[d, f]}{(1 + \frac{L}{U})\widehat{\mathrm{Var}}[f]}$$

Compute

$$\widehat{\delta_{1,2}^{PP}} \leftarrow \lambda^\star \bar{f}_U + \frac{1}{L}\sum_{j=1}^{L}(d_j - \lambda^\star f_j)$$

Compute

$$\widehat{\mathrm{Var}}[\widehat{\delta_{1,2}^{PP}}] \leftarrow \frac{1}{L}\widehat{\mathrm{Var}}[d] - \frac{U\left(\widehat{\mathrm{Cov}}[d, f]\right)^2}{L(L+U)\widehat{\mathrm{Var}}[f]}$$

Compute

$$p_{1,2}^{PP} \leftarrow 1 - \Phi\left(\frac{\widehat{\delta_{1,2}^{PP}}}{\sqrt{\widehat{\mathrm{Var}}[\widehat{\delta_{1,2}^{PP}}]}}\right)$$

return $p_{1,2}^{PP}$

Algorithm 4 Human-only Paired Permutation Test

Input: human scores $\{Y_{1j}, Y_{2j}\}_{j=1}^{L}$, number of permutations B
Output: one-sided p-value for testing $H_0 : \delta_{1,2} \leq 0$ vs. $H_1 : \delta_{1,2} > 0$
Compute

$$\widehat{\delta_{1,2}^{H}} \leftarrow \frac{1}{L}\sum_{j=1}^{L}(Y_{1j} - Y_{2j})$$

for $b = 1, \ldots, B$ **do**
  Draw $\varepsilon_j^{(b)} \overset{\text{iid}}{\sim} \mathrm{Uniform}\{-1, +1\}$ for $j = 1, \ldots, L$
  Compute

$$\widehat{\delta_{1,2}^{H,(b)}} \leftarrow \frac{1}{L}\sum_{j=1}^{L}\varepsilon_j^{(b)}(Y_{1j} - Y_{2j})$$

end for
Compute

$$p_{1,2}^{H} \leftarrow \frac{1 + \sum_{b=1}^{B}\mathbb{1}\left\{\widehat{\delta_{1,2}^{H,(b)}} \geq \widehat{\delta_{1,2}^{H}}\right\}}{B+1}$$

return $p_{1,2}^{H}$

Algorithm 5 Auto-only Paired Permutation Test

Input: metric scores $\{F(X_{1j}), F(X_{2j})\}_{j=1}^{L}$, number of permutations B
Output: one-sided p-value for testing $H_0 : \delta_{1,2} \leq 0$ vs. $H_1 : \delta_{1,2} > 0$
Compute

$$\widehat{\delta_{1,2}^{A}} \leftarrow \frac{1}{L}\sum_{j=1}^{L}(F(X_{1j}) - F(X_{2j}))$$

for $b = 1, \ldots, B$ **do**
  Draw $\varepsilon_j^{(b)} \overset{\text{iid}}{\sim} \mathrm{Uniform}\{-1, +1\}$ for $j = 1, \ldots, L$
  Compute

$$\widehat{\delta_{1,2}^{A,(b)}} \leftarrow \frac{1}{L}\sum_{j=1}^{L}\varepsilon_j^{(b)}(F(X_{1j}) - F(X_{2j}))$$

end for
Compute

$$p_{1,2}^{A} \leftarrow \frac{1 + \sum_{b=1}^{B}\mathbb{1}\left\{\widehat{\delta_{1,2}^{A,(b)}} \geq \widehat{\delta_{1,2}^{A}}\right\}}{B+1}$$

return $p_{1,2}^{A}$

Algorithm 6 Prediction-powered Paired Permutation Test

1: **Input:** system outputs and their human scores $\{X_{1j}, X_{2j}, Y_{1j}, Y_{2j}\}_{j=1}^{L}$, system outputs without human scores $\{X_{1j}, X_{2j}\}_{j=L+1}^{L+U}$, automatic metric F, number of permutations B
2: **Output:** one-sided p-value for testing $H_0 : \delta_{1,2} \leq 0$ vs. $H_1 : \delta_{1,2} > 0$
3: Compute $M \leftarrow L + U$
4: Set $d_j \leftarrow Y_{1j} - Y_{2j}$ for $j = 1, \ldots, L$; $f_j \leftarrow F(X_{1j}) - F(X_{2j})$ for $j = 1, \ldots, M$
5: Compute $\widehat{\delta_{1,2}^{PP}}$ from $\{d_j\}_{j=1}^{L}, \{f_j\}_{j=1}^{M}$ via steps 4–11 of Alg. 3
6: **for** $b = 1, \ldots, B$ **do**
7: Draw $\varepsilon_j^{(b)} \overset{\text{iid}}{\sim} \text{Uniform}\{-1, +1\}$ for $j = 1, \ldots, M$
8: Set $d_j^{(b)} \leftarrow \varepsilon_j^{(b)} d_j$ for $j = 1, \ldots, L$
9: Set $f_j^{(b)} \leftarrow \varepsilon_j^{(b)} f_j$ for $j = 1, \ldots, M$
10: Compute $\widehat{\delta_{1,2}^{PP,(b)}}$ from $\{d_j^{(b)}\}_{j=1}^{L}, \{f_j^{(b)}\}_{j=1}^{M}$ via steps 4–11 of Alg. 3 (re-estimating $\lambda^\star$)
11: **end for**
12: Compute

$$p_{1,2}^{PP} \leftarrow \frac{1 + \sum_{b=1}^{B} \mathbb{1}\left\{\widehat{\delta_{1,2}^{PP,(b)}} \geq \widehat{\delta_{1,2}^{PP}}\right\}}{B+1}$$

13: **return** $p_{1,2}^{PP}$